UNIVERSIDADE FEDERAL DA PARAÍBA

DISCIPLINA:
Estatística e Planejamento - Experimentos

Prof: Dr Luiz Bueno da Silva

Aluno: Edgar Massaru Yoshida

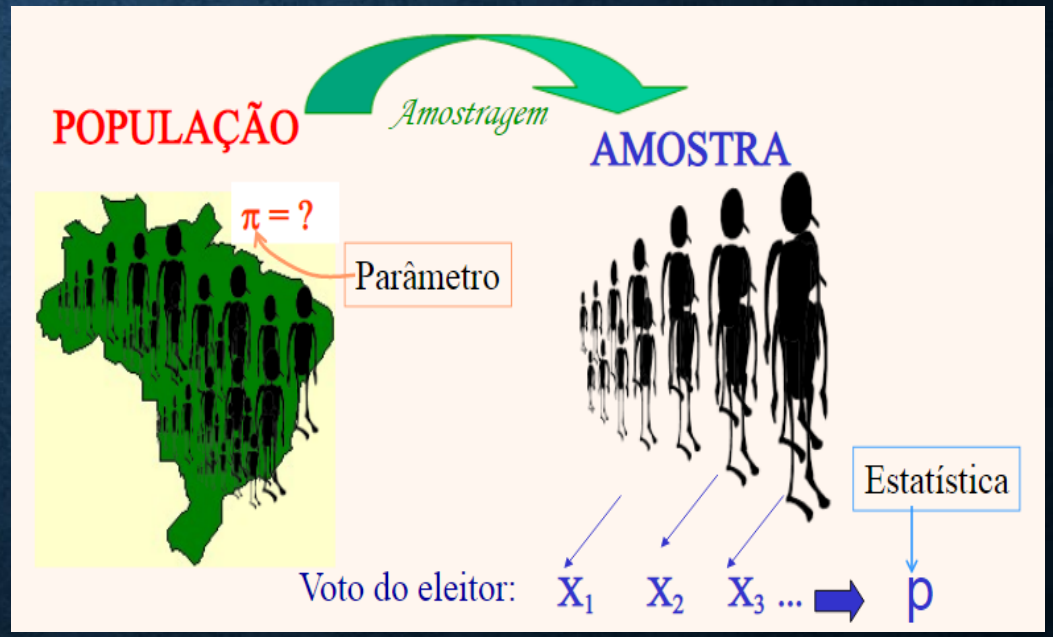
UNIVERSIDADE FEDERAL DA PARAÍBA

Artigo:

"How big my sample need to be? A primer on the murky world of sample size estimation."

Autores: Batterham, A.M.; Atkinson, G.

Publicação: Physical Therapy in Sport – 2005, p. 153-163



CONTEXTUALIZAÇÃO

- O objetivo deste artigo é levantar algumas questões sobre a importância de se fazer um planejamento da quantidade de amostras em pesquisas, e que os métodos aplicados para isso devem ser coerentes com o grau de precisão requerida nos resultados pesquisados.

- Ele não leva a fórmulas prontas de determinação das quantidades, mas o caminho e os cuidados a se tomar para se escolher e usar as teorias e métodos estatísticos existentes.

METODOLOGIA

- Para mostrar os cuidados a serem tomados, foi adotado um modelo hipotético de grupo de controle e outro de tratamento, para variar determinados parâmetros a fim de visualizar os seus impactos na quantidade de amostras necessárias.

- É baseado na tradicional teste “t” da hipótese nula, estimando o tamanho da amostra dentro do nível de significância que se espera, e da potência do teste, seguindo os princípios gerais e conceitos consagrados, para que possa ser feita a generalização do estudo.

REVISÃO DE CONCEITOS

p-value e o erro Tipo I:

- p-value menor que o nível de significância considerado, rejeita-se a hipótese nula. É o chamado erro "Type-I" ou "false-alarm" (alpha). Quando rejeito a hipótese de que $H_0 = 0$ é verdadeiro.

Potência do teste e o erro Tipo II:

- A verificação do erro "Type-II" se trata da falha em se rejeitar a hipótese nula, ou "failed-alarm" (beta). Quando rejeito a hipótese de que $H_0 = 0$ é falso.
- A potência do teste é o valor do complemento de "beta" para 100%.

REVISÃO DE CONCEITOS

- p-value e o erro Tipo I
- Potência do teste e o erro Tipo II

Decisão	H_0 é verdadeiro	H_0 é falso
Aceito H_0	Decisão Correta (probabilidade = 1 - α)	Erro tipo II - deixa de rejeitar H_0 quando ela é falsa (probabilidade = β)
Rejeito H_0	Erro tipo I - rejeitando H_0 quando ele é verdadeiro (probabilidade = α)	Decisão Correta (probabilidade = 1 - β)

REVISÃO DE CONCEITOS

Para determinação de amostras, os autores fazem as seguintes ressalvas:

- Convencionalmente é aceito o valor de “alpha”= 5% e de “beta”= 10%, na maioria dos casos.
- Para estimativa do tamanho da amostra é necessário a verificação do erro “Type-II” .
- Quanto menor o valor de “alpha”, maior a probabilidade de se incorrer no erro “Type-II”.
- Caso o estudo envolva situações que possam resultar em riscos grandes, morte por exemplo, os valores de “alpha” e “beta” devem ser menores, “alpha”= 1% por exemplo.
- E para situações onde o resultado sejam situações sem riscos expressivos, pode-se assumir “beta”= 20% por exemplo.

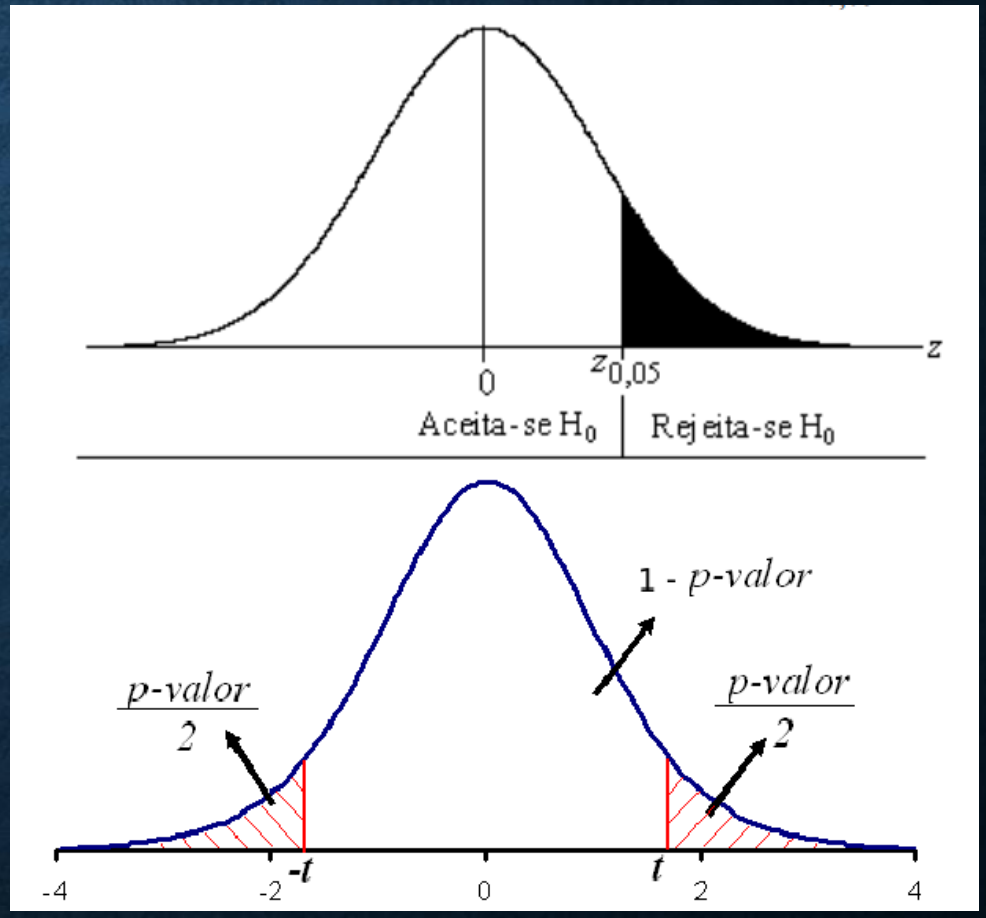
REVISÃO DE CONCEITOS

One tailed (teste de um dos extremos):

- quando a variação se dá numa direção, por exemplo, podemos assumir a hipótese de que o decréscimo da dor se dá apenas por determinado tratamento.

Two tailed (teste dos dois extremos):

- quando a hipótese se centra na mudança, não importando a direção.



REVISÃO DE CONCEITOS

Para determinação de amostras, os autores fazem a seguinte ressalva:

- Na grande maioria dos casos o procedimento indicado é o “two-tailed” (Altman, 1991) mesmo quando se tem uma grande convicção de que o tratamento só poderia seguir numa direção, pois ainda assim sempre haverá a possibilidade que outros fatores afetem o resultado avaliado.

APRESENTAÇÃO DO EXPERIMENTO

- Dois grupos de amostras: grupo tratado e grupo controle.
- Comparação entre o grupo tratado e de grupo de controle.
- Variável: mudanças das médias do grau de dor.
  - Os valores obtidos por medida contínua, ou por escala visual em categorias.
- Mesma quantidade de amostras nos dois grupos.
- Dados do Grupo tratado e resultado na Coluna A da tabela a seguir.
- Variações de parâmetros do grupo tratado e verificação dos efeitos no número de amostras (Colunas de B até F).

RESULTADOS DO EXPERIMENTO

- Usando o teste de hipótese t com nível de significância de 5%, e potência do teste de 90% de detectar uma diferença na média de dor a cada unidade de variação dessa escala, chegou-se a amostra com 23 indivíduos em cada grupo, assumindo que o desvio padrão seja também de 1 unidade. **Coluna A** da tabela abaixo:

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (*d*/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

RESULTADOS DO EXPERIMENTO

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (*d*/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

- Na **coluna B**, se alterou o desvio padrão para 0,45 assumindo que a correlação das média seja de r=0,9, e mantendo-se iguais os demais fatores. O chamado "effect size", dado como a diferença entre as médias dividido pelo desvio padrão comum, passou para 2,22 que se traduz numa amostra menor, de n=5 ao invés dos 23 obtidos na coluna A.

RESULTADOS DO EXPERIMENTO

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (d/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

- Na **coluna C**, os autores alteraram o procedimento de "two-tailed" para "one-tailed" por saber que neste caso específico ele é mais apropriado (Knottnerus & Bouter, 2001), mantendo todas as demais condições da coluna A, obtendo-se assim um n=18, menor que os 23 obtidos no procedimento "two-tailed".

RESULTADOS DO EXPERIMENTO

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (*d*/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

- Na **coluna D**, foi dobrada a diferença entre as médias dos grupos, e se mantiveram as demais condições da coluna A. Verifica-se que o valor mínimo do Intervalo de Confiança aumentou em relação a condição da coluna A, o que significa que a diferença da média entre amostras é maior que 1. Houve significativa diminuição do número de amostras (n= 7 ao invés de 23), e o "effect size" dobrou pois foi mantida a mesma variabilidade.

RESULTADOS DO EXPERIMENTO

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (d/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

- Na **coluna E**, foi dobrado o desvio padrão (e portanto, atestando maior variabilidade), o que em outras palavras, se aumentou o ruído nos dados. E quanto mais ruído na amostra, mais dados são necessários para se obter sinais significativos de diferenças de médias. Neste caso o número de amostras por grupo saltou para 86 (na coluna A eram 23).

RESULTADOS DO EXPERIMENTO

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_d$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (*d*/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

- Na **coluna F**, foi diminuído a potência do teste para 80%, mantendo-se os demais valores dos parâmetros da coluna A. Diminuir a potência do teste, significa aumentar a aceitação de uma maior probabilidade de ocorrer o erro tipo II (“beta”) em relação a coluna A. Como resultado se necessitará de menos amostras já que se aceita um erro maior, e assim passou-se para n= 17 ao invés de 23.

OBSERVAÇÕES DOS AUTORES

- Essas análises foram feitas para amostras de mesmo tamanho, o que leva a uma redução da variabilidade dos dados (Atkinson & Nevill, 2001).
- Nem sempre isso é possível se ter o mesmo número de amostras, e nesses casos, pode-se utilizar essas variáveis introduzindo-as como covariantes nas análises. (Vickers & Altman, 2001).
- Contudo, uma pesquisa com quantidade de amostras desbalanceadas, leva a necessidade de uma quantidade de amostras total maior.
- Para mesma quantidade total de amostras, a situação de termos mais amostras num grupo que no outro, leva a resultados estatísticos mais pobres do que se os dois grupos tivessem a mesma quantidade. (Whitley & Ball, 2002).

DETERMINAÇÃO DO NÚMERO DE AMOSTRAS
PASSO A PASSO

- Responder a pergunta: Vale a pena realizar este estudo? Caso a resposta seja "não é", é necessário que se analise e discuta antes de prosseguir com ele. Caso seja afirmativo, seguem-se os itens na sequência.
- Determinar as bases de comparações, por exemplo, quais as características dos componentes e a forma que serão agrupados (tipo de casualização), tanto do grupo de controle como do grupo tratado, e quais serão as intervenções tanto do grupo controle como nos grupos tratados.
- Identificar a variável independente primária (o resultado). É importante saber quando serão medidos, se será logo após o tratamento, ou se após um período de acompanhamento pós tratamento.
- Identificar as variáveis dependentes (os tratamentos). É importante saber qual os limites mínimo e máximo a serem considerados.

Até aqui os autores chamam de itens básicos que compõem o PICO (Participants, Interventions, Comparisons, and Outcomes).

DETERMINAÇÃO DO NÚMERO DE AMOSTRAS
PASSO A PASSO

- Identificar qual é o menor efeito a ser considerado dentro da escala de efeitos. Pode-se também identificar na literatura existente sobre o tema, qual a variação mínima entre as médias que deverá ser considerada. E porque adotá-la? Na falta de literaturas sobre o tema, conduzir estudos preliminares com menos amostras afim de obter uma estimação dessa variação mínima a ser considerada, e compartilhar esses estudos com outros pesquisadores para validar essa estimação. Diferença entre médias estimada= "μ".
- Identificar na literatura existente sobre o tema, qual o desvio padrão das médias a ser considerado. Desvio Padrão= "sd".
- Decidir qual o risco a ser assumido na tomada de decisões: Erro tipo I ou Erro tipo II? Teste de significância para One ou Two-sided? Existe algum motivo forte para se testar apenas One-sided?
- Assumir o nível de significância "alpha" entre 1% e 5%, e a potência do teste "1- beta" entre 80% a 90%. Os autores recomendam utilizar "alpha"= 5% e "beta"= 10%.

DETERMINAÇÃO DO NÚMERO DE AMOSTRAS
PASSO A PASSO

- Há algum motivo para que os grupos não possam ter a mesma quantidade de amostras?
- Caso haja diferenças consideráveis entre as médias, usa-se o teste estatístico "t" de forma independente dentro de um intervalo de confiança, e caso não haja, pode-se utilizar o teste "t" entre grupos, contudo há a alternativa de se utilizar a análise da covariância. Neste caso se assume como variável dependente os resultados do tratamento, o grupo nominal como variáveis independentes, e os valores antes do tratamento como covariantes. Este tipo de análise também auxilia na identificação de perdas que se traduzam em resultados inexpressivos causados devido ao tempo entre o tratamento e a tomada desses resultados. Devido a isso, em geral, se estima acrescer de forma arbitrária em 10% o número de amostras para que essa perda seja compensada.
- Para estimar a quantidade da amostra, escolher entre usar um software feito para isso, ou utilizar nomograma.

SUGESTÕES DOS AUTORES
SOFTWARES

- nQuery Advisor (Statistical Solutions, Cork, EIRE)
- Stata
- SAS
- StatsDirect

SUGESTÕES DOS AUTORES
NOMOGRAMA

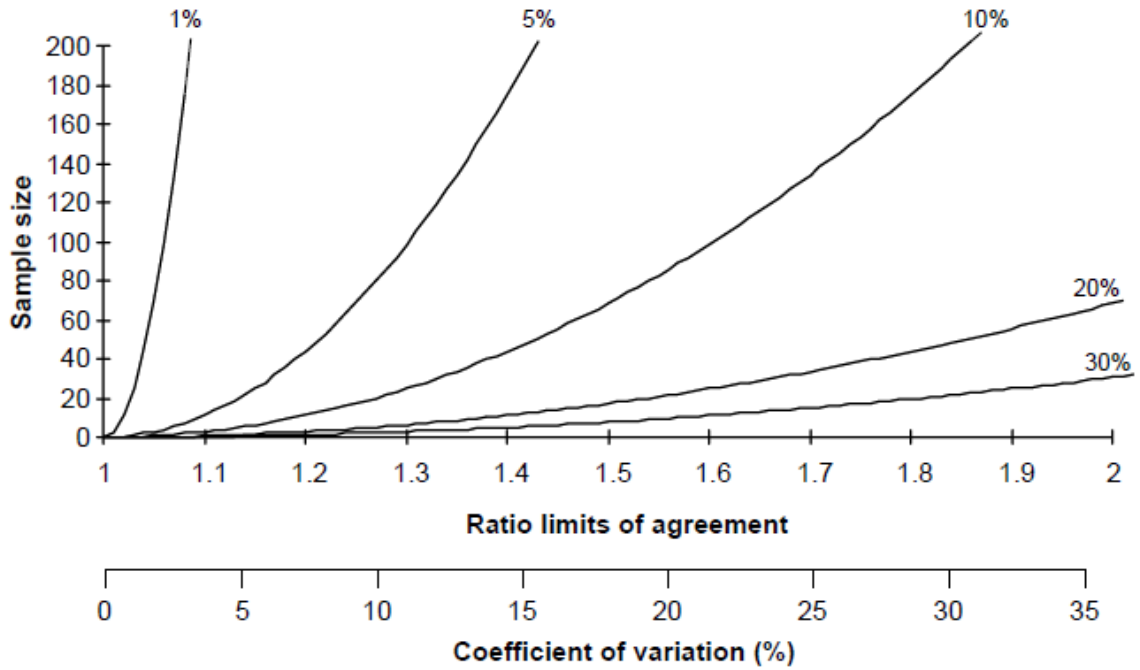


Fig. 1. A nomogram to estimate the effects of measurement repeatability error on whether 'analytical goals' are attainable or not in exercise physiology research. Statistical power is 90%. The different lines represent different worthwhile changes of 1, 5, 10, 20 and 30% due to some hypothetical intervention. The measurement error statistics, which can be utilised are the LOA and CV. For example, a physiological measurement tool, which has a repeatability CV of 5% would allow detection of a 5% change in a pre-post design experiment (using a paired *t*-test for analysis of data) and with a feasible sample size (approximately 20 participants).

REFERÊNCIAS DO TEXTO

- Altman, D. G. (1991). Practical statistics for medical research. London: Chapman & Hall.
- Atkinson, G., & Nevill, A. M. (2001). Selected issues in the design and analysis of sport, performance research. Journal of Sports Sciences, 19, 811–827.
- Knottnerus, J. A., & Bouter, L. M. (2001). The ethics of sample size: Twosided testing and one-sided thinking. Journal of Clinical Epidemiology, 54, 109–110.
- Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. British Medical Journal, 323, 1123–1124.
- Whitley, E., & Ball, J. (2002). Statistics review 4: Sample size calculations. Critical Care, 6, 335–341.