Projeto final ME-110
Entrega: 26/06/2017

- O projeto final constirá de 2 partes.
- Este trabalho final pode ser realizado em equipes de até 4 pessoas. Os alunos com notas menores que 5 na P1 somente poderão fazer grupo com alunos com notas não mais do que 30 pontos aparte. Somente um relatório deve ser entregue (por favor, não se esqueçam de colocar todos os nomes dos membros da equipe no relatório). A equipe deve se manter a mesma para as duas partes do trabalho final.
- Os relatórios devem ser submetidos de duas formas: eletronicamente (arquivo base TrabalhoFinal.Rmd e anexos necessários) e impresso (.pdf gerado a partir do .Rmd). Por favor, façam relatórios legíveis e caprichados. As notas serão baseadas em sua criatividade, adequabilidade das técnicas estatísticas, interpretação dos resultados e apresentação. Uso eficiente de cores, gráficos e ilustrações engrandecem um relatório.

Parte 1: Mega Sena

Quando analisamos os dados da Mega-Sena, cada retirada não pode ser considerada um ensaio multinomial pois a cada sorteio temos 6 bolas retiradas **sem** reposição. Sendo assim, usar o teste qui-quadrado de bondade de ajuste não é a análise correta.

Para analisar estes dados vamos usar um procedimento para calcular o p-valor baseado em simulações.

Continuamos com a hipótese nula de que todos os resultados são igualmente prováveis. Para verificar isto escolha uma amostra aleatória simples de $N = 1000$ sorteios da Mega-Sena (seja bastante explicito em seu trabalho como foram escolhidos os sorteios). Neste caso, como em cada sorteio seleciona-se 6 bolas, sua amostra consiste em $n = 6N$ observações (não independentes!).

Utilize como estatística do teste:

$$W = \sum_{j=1}^{60} |O_j - E_j|,$$

onde $E_j = n/60$.

Use os seus dados para calcular o valor observado da estatística do teste w^*_{obs}.

O problema é que neste caso não sabemos a distribuição aproximada de W nem mesmo sob a hipótese nula.

Simulação:

- Cada rodada l (*run*) consistirá em N retiradas de 6 números do conjunto $\{1, 2, \ldots, 60\}$ sem reposição.
- Cada rodada apresentará $n_l = 6N$ números.
- Usar estes n_l números para calcular w^*_l.
- Repetir o procedimento M vezes ($l = 1, \ldots, M$).
- Contar quantas vezes obteve-se $w^*_l > w^*_{obs}$ ($= x_{simul}$).
- Usar $\hat{p}_{simul} = x_{simul}/M$ como estimativa para o p-valor.
- Construir um IC de nível $(1-\alpha)100\%$ para o p-valor considerando que cada rodada é um ensaio de Bernoulli.

Utilize um histograma para apresentar o resultado das simulações. Indique no histograma o valor observado da estatística.

Apresente a estimativa para o p-valor.

Com base na simulação acima chegar à conclusão se todos os números da Mega-Sena tem a mesma chance de ocorrer.

Parte 2: Testes de permutação: Excolha um dos dois problemas abaixo:

1.- Teste de Fisher para mais do que dois tratamentos: Vimos em classe que no teste de homogeneidade, a aproximação através da distribuição quiquadrado somente pode ser aplicada quando todos os valores esperados são maiores que 5. No caso de experimentos completamente aleatorizados e tabelas 2x2 podemos aplicar o teste exato de Fisher. O objetivo desta parte é realizar o teste exato de Fisher para experimentos completamente aleatorizados com 3 tratamentos e duas respostas. Neste caso, não é possível realizar o teste exato e o p-valor pode ser obtido através de simulação.

Realize um experimento completamente aleatorizado (CRD) balanceado com 3 tratamentos e duas respostas para estudar algum assunto de seu interesse. Escolha 30 unidades experimentais e aloque 10 unidades experimentais para cada tratamento.

Seu relatório, além de analizar completamente o problema, deve responder as seguintes questões:

- Por que este assunto te interessa?
- Qual é o seu estudo?
- Alguma coisa surpreendente aconteceu enquanto da coleta de dados?
- Verificação de todas as suposições feitas.

A seguir, seu relatório deve apresentar e resumir seus dados usando tabelas e gráfixos. Realize um teste de Fisher para a alternativa de que as populações não são homogêneas. Utilize como estatística do teste:

$$W = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

onde $E_{ij} = n_i m_j / n$. O problema é que neste caso **não** sabemos a distribuição aproximada de W nem mesmo sob a hipótese nula. (**NÃO** utilize a aproximação qui-quadrado, você deve encontrar esta distribuição através de simulações).

Simulação:

- Cada rodada l (*run*) consistirá em em montar uma tabela de contingência 3 x 2 mantendo as marginais fixas.

- Cada rodada l apresentará uma tabela.
- Usar esta tabela para calcular w_l^*.
- Repetir o procedimento M vezes ($l = 1, \dots, M$).
- Contar quantas vezes obteve-se $w_l^* > w_{obs}^*$ ($= x_{simul}$).
- Usar $\hat{p}_{simul} = x_{simul}/M$ como estimativa para o p-valor.
- Construir um IC de nível $(1-\alpha)100\%$ para o p-valor considerando que cada rodada é um ensaio de Bernoulli.

2.- Teste de permutação para comparação de duas médias:
Realize um experimento completamente aleatorizado (CRD) balanceado com 2 tratamentos e uma reposta numérica. Escolha 20 unidades experimentais e aloque 10 unidades experimentais para cada tratamento.

Seu relatório, além de analizar completamente o problema, deve responder as seguintes questões:

- Por que este assunto te interessa?
- Qual é o seu estudo?
- Alguma coisa surpreendente aconteceu enquanto da coleta de dados?
- Verificação de todas as suposições feitas.

A seguir, seu relatório deve apresentar e resumir seus dados usando gráficos. Realize um teste de permutação para testar se a média das duas sub-populações é a mesma, contra a alternativa de que não são as mesmas. Utilize como estatística do teste:

$$W = |\bar{X} - \bar{Y}|$$

onde $\bar{X}$ e $\bar{Y}$ são as médias amostras das respostas dos indivíduos alocados ao tratamento 1 e tratamento 2 respectivamente. O problema é que neste caso **não** sabemos a distribuição aproximada de W nem mesmo sob a hipótese nula. (NÃO utilize a aproximação normal para a distribuição da estatística do teste sob H0, você deve encontrar esta distribuição através de simulações.)

Simulação:

- Cada rodada l (*run*) consistirá em permutar os 20 valores mantendo dois grupos de 10 respostas.
- Cada rodada l apresentará duas médias $\bar{X}_l$ e $\bar{Y}_l$.
- Usar estas médias para calcular w_l^*.
- Repetir o procedimento M vezes ($l = 1, \ldots, M$).
- Contar quantas vezes obteve-se $w_l^* > w_{obs}^*$ ($= x_{simul}$) (ou a estatística e a decisão adequados se usar outra hipótese alternativa).
- Usar $\hat{p}_{simul} = x_{simul}/M$ como estimativa para o p-valor.
- Construir um IC de nível $(1-\alpha)100\%$ para o p-valor considerando que cada rodada é um ensaio de Bernoulli.