# Análise de Dados em Astronomia I: AGA 0505

Eduardo Cypriano

IAG/USP

Primeiro semestre de 2017

1. Índice

2. W&J Cap 1: O processo de decisão em ciência
   - Como a ciência é feita?
   - Probabilidades e distribuições probabilísticas
   - Tudo já resolvido?
   - Probabilidade e estatística em inferência: uma visão geral desse curso

# Informações Gerais

Objetivo do curso: ensinar noções de práticas de estatística aplicada a análise de dados, com enfoque (levemente) Bayesiano e voltada para astrônomos

Referência básica:
**Practical Statistics for Astronomers.** (Wall & Jenkins, 2012 - Segunda edição)

# Informações Gerais

- Aulas segundas-feiras das 14:00 às 16:00 - Sala A-304

- Data importantes
  - Primeira Prova: 11/04
  - Segunda Prova: 30/05
  - Substitutiva: 11/06 (Aberta: todos podem fazer e pode substituir ou não uma das notas anteriores)

- Nota Final: Cada prova vale 35% da nota. Exercícios e listas valem 30%.

# Informações Gerais

- O curso será fortemente baseado no **Practical Statistics for Astronomers**. Deveremos ver até o capítulo 6 até o final do curso (e algo do cap. 7 se houver tempo). Aproximadamente dos capítulos 1 ao 4 até a primeira prova e os capítulos 5 e 6 até a segunda.
- O professor irá produzir notas de aula que são apenas para dar o "fio da meada" durante as aulas.
- Ao longo do curso passarei passarei problemas e exercícios . Peço atenção ao prazos de entrega (geralmente para a próxima aula até as 14:30) pois esses exercícios serão fundamentais para consolidação dos conceitos que veremos em aula.
- Os estudantes podem usar a linguagem de computação que quiserem para fazer os exercícios (ainda que algumas funções de alto nível serão proibidas em determinadas situações).
- O professor irá usar a linguagem R para dar os exemplos em sala de aula e recomenda a qualquer estudante mais interessado em estatística e análise de dados a ser informar a respeito dessa linguagem em http://www.r-project.org/.

# Informações Gerais

- O curso tem um monitor, André Zamorano Vitorelli (IAG/USP - andrezvitorelli at gmail.com), que poderá auxilia-los com os trabalhos. Ele poderá ser localizado na sala B-305.

- O curso fará use extensivo do MOODLE/STOA: `http://disciplinas.stoa.usp.br/course/view.php?id=12313`. As listas, por exemplo, poderão ser encontradas lá.

- A forma recomendada para o estudante tirar uma dúvida é posta-la do *Fórum de dúvidas*. Dessa forma todos os colegas poderão se beneficiar das respostas.

- Se por qualquer razão a pergunta não puder ser postada no fórum o aluno pode procurar o monitor e o professo via e-mail ou pessoalmente.

# Objetivo do Curso

- Esse é um curso essencialmente prático (notem o nome da bibliografia principal).
- O objetivo aqui é familiarizar o estudante com conceitos de estatística sem uma preocupação com provas rigorosas.
- Serão apresentadas algumas técnicas de análise de dados que servirão para que os estudantes tomem conhecimento de sua existência e aprenda a aplicá-las.
- A ideia é que o aluno ganhe independência para procurar a saber usar a ferramenta para a análise que se apresentar no seu caminho.

# Sugestão de bibliografia auxiliar

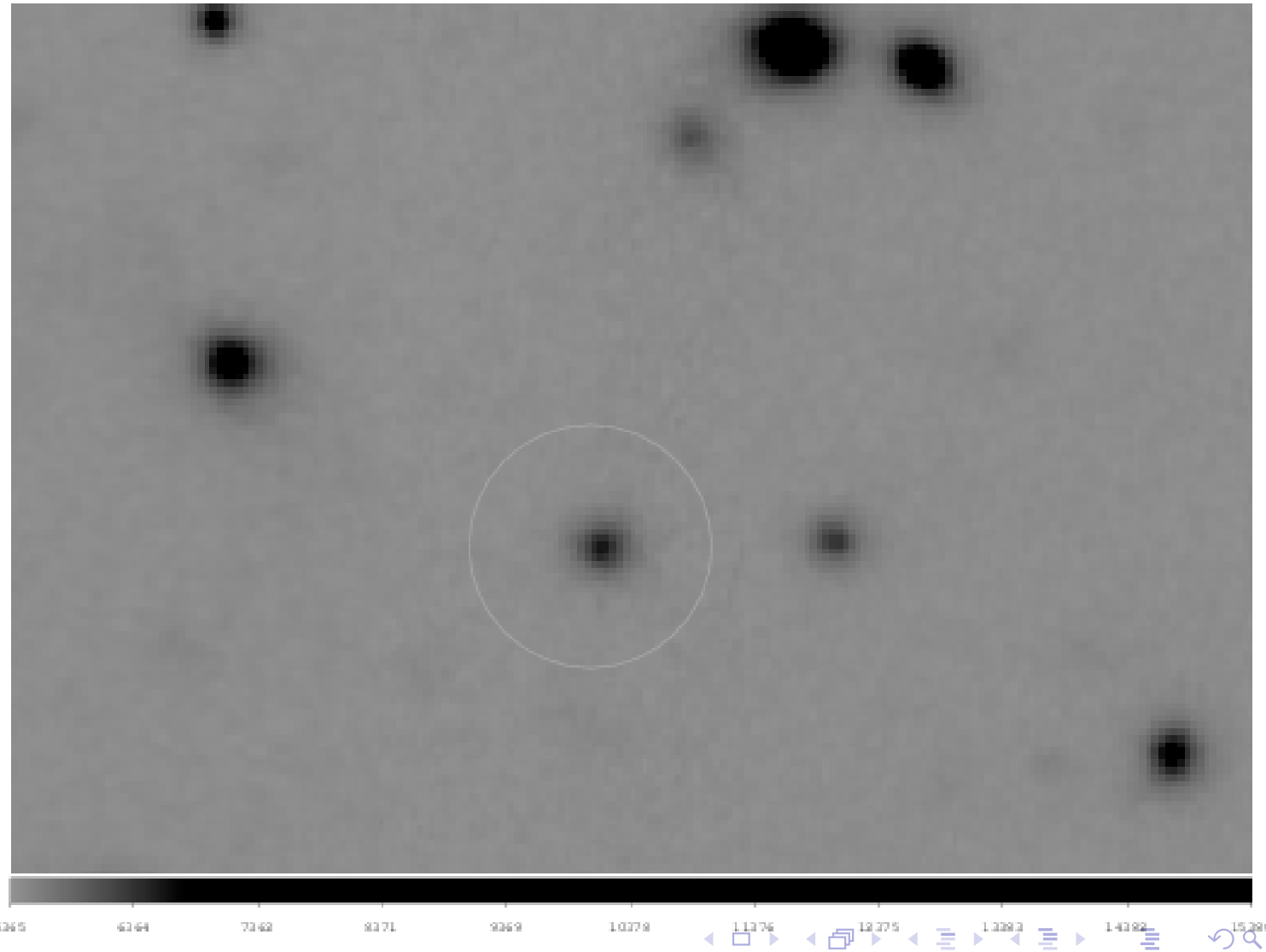
- *Modern Statistical Methods for Astronomy: With R Applications* (Eric D. Feigelson & G. Jogesh Babu 2012) - Conteúdo similar
- *Numerical Recipes: The Art of Scientific Computing* (Press et al., 2007) - Implmentação prática dos métodos
- *Data Reduction and Error Analysys for the Physical Sciences* (Beavington & Robinson 2002) – "Bíblia" do frequentismo
- *Data Analysis: A Bayesian Tutorial* (Sivia, 2006) - Introdução à estatística Bayesiana

# W&J Cap 1: O processo de decisão em ciência

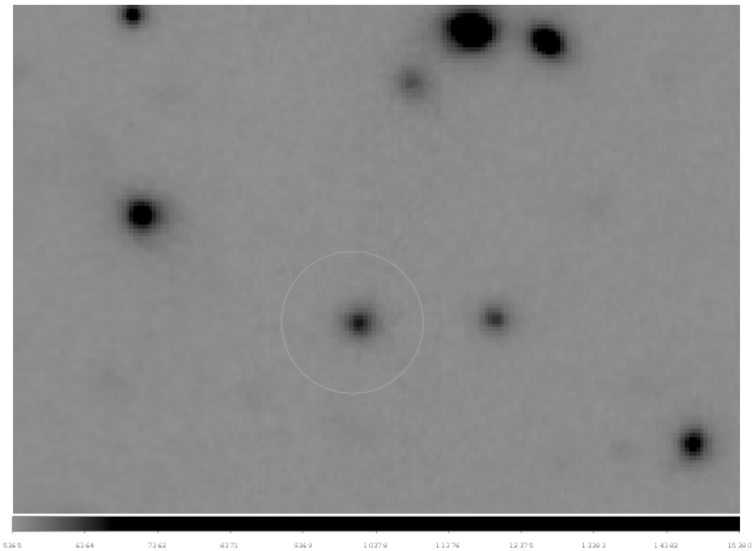
- Tomar decisões (fundamentadas) é essencial para a ciência.
  - Essa hipótese ou teoria está correta? Se não, por que não?
  - Esses dados são auto-consistentes? São consistentes com outros dados? São adequados para responder à questão colocada?
  - ...

# W&J Cap 1: O processo de decisão em ciência

- Decisões são tomadas frequentemente por comparação.
  - Esse borrãozinho na imagem é uma estrela ou uma galáxia?



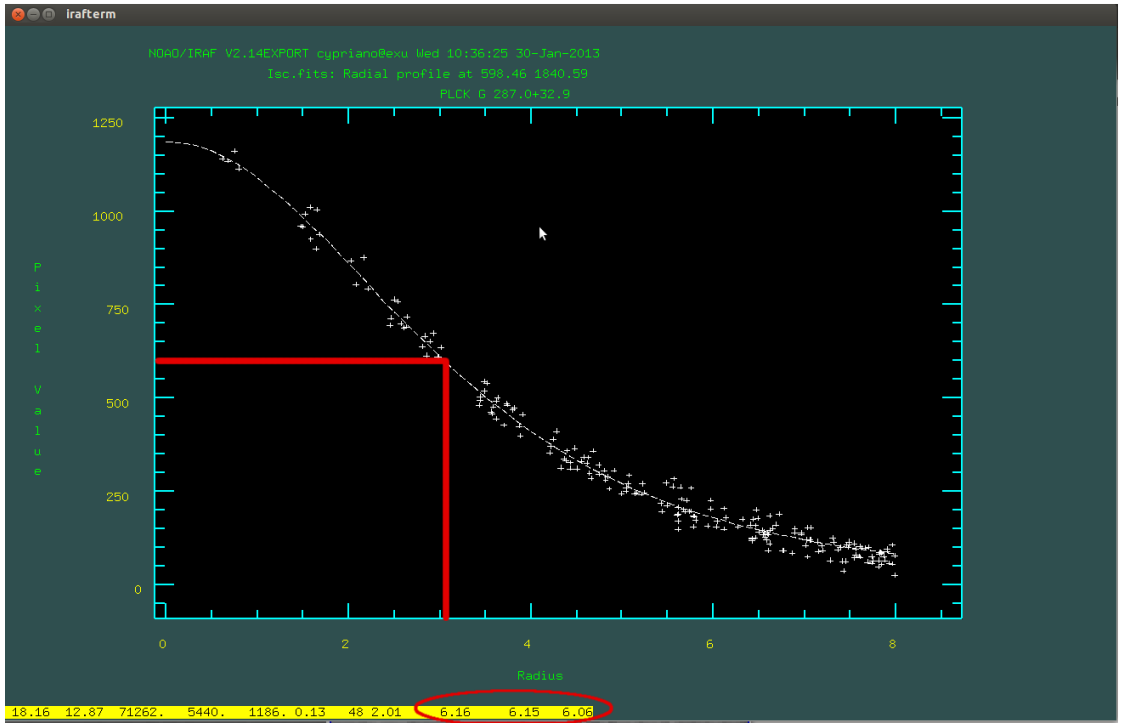
# W&J Cap 1: O processo de decisão em ciência



- Uma caracterização primária da imagem é através da sua FWHM (Full Widht at Half Maximum ou largura a meia-altura)

- Ao olhar a posição do valor da FWHM desse objeto dentro da distribuição da FWHM de todos os objetos pode ajudar na tomada de decisão.

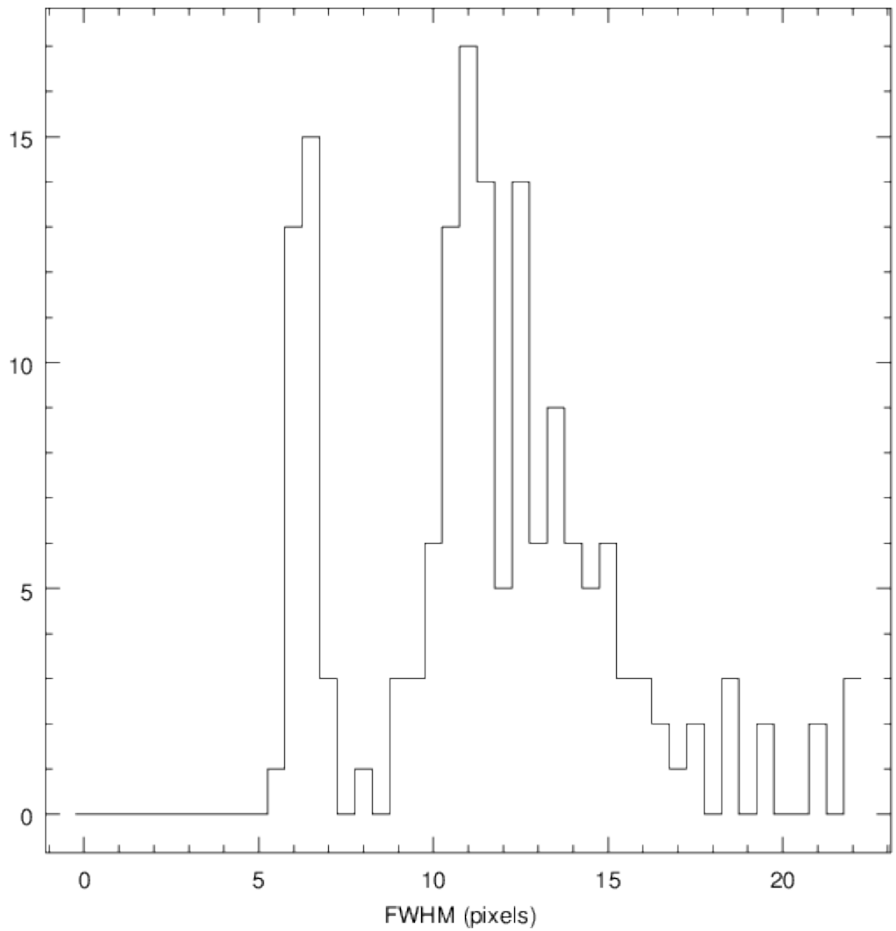
- Na figura abaixo pode ser visto o perfil de luz desse objeto, ou seja o valor da cada pixel em função à distância ao centro do objeto.




- A linha vermelha horizontal representa a "meia altura", ou seja a intensidade máxima dividida por dois.
- A linha vermelha vertical indica a que distância do centro o perfil de luz cai até esse valor, no caso algo por volta de 3 pixeis. Essa é metade da largura a meia altura, já que nessa figura os raios tem valores positivos apenas. [1]

[1] Os números da linha de baixo em destaque (6.16, 6.15, 6.06), indicam os valores estimados pelo IRAF/IMEXAM para a FWHM desse objeto.

- Na figura abaixo pode ser visto um histograma que mostra a distribuição dos valores da FWHM para a a imagem que contém o objeto em questão.




- É sabido que objetos não-resolvidos (e.g. estrelas) tem uma valor aprox. constante da FWHM e que não há valores (corretos) menores do que estes.
- Assim sendo, o nosso objeto é uma estrela ou galáxia?

- Nesse processo a imagem do objeto foi descrita por uma *estatística* (a FWHM) e isso permitiu que a decisão fosse tomada.

- **Estatísticas** permitem a tomada de decisões (inferência estatística) porque temos parâmetros de comparação.

- Toda medida que fazemos ou parâmetro ou valor que determinamos deve ser acompanhado de uma *estimativa de erro* ou *incerteza*

- Trata-se de uma um intervalo (dado em termos de probabilidade) que engloba nossa crença no valor verdadeiro dessa quantidade.

- Qualquer quantidade medida ou estimativa tem pouca ou nenhuma valia para a tomada de uma decisão, ou seja, para ciência, a menos que acompanhada de sua incerteza.

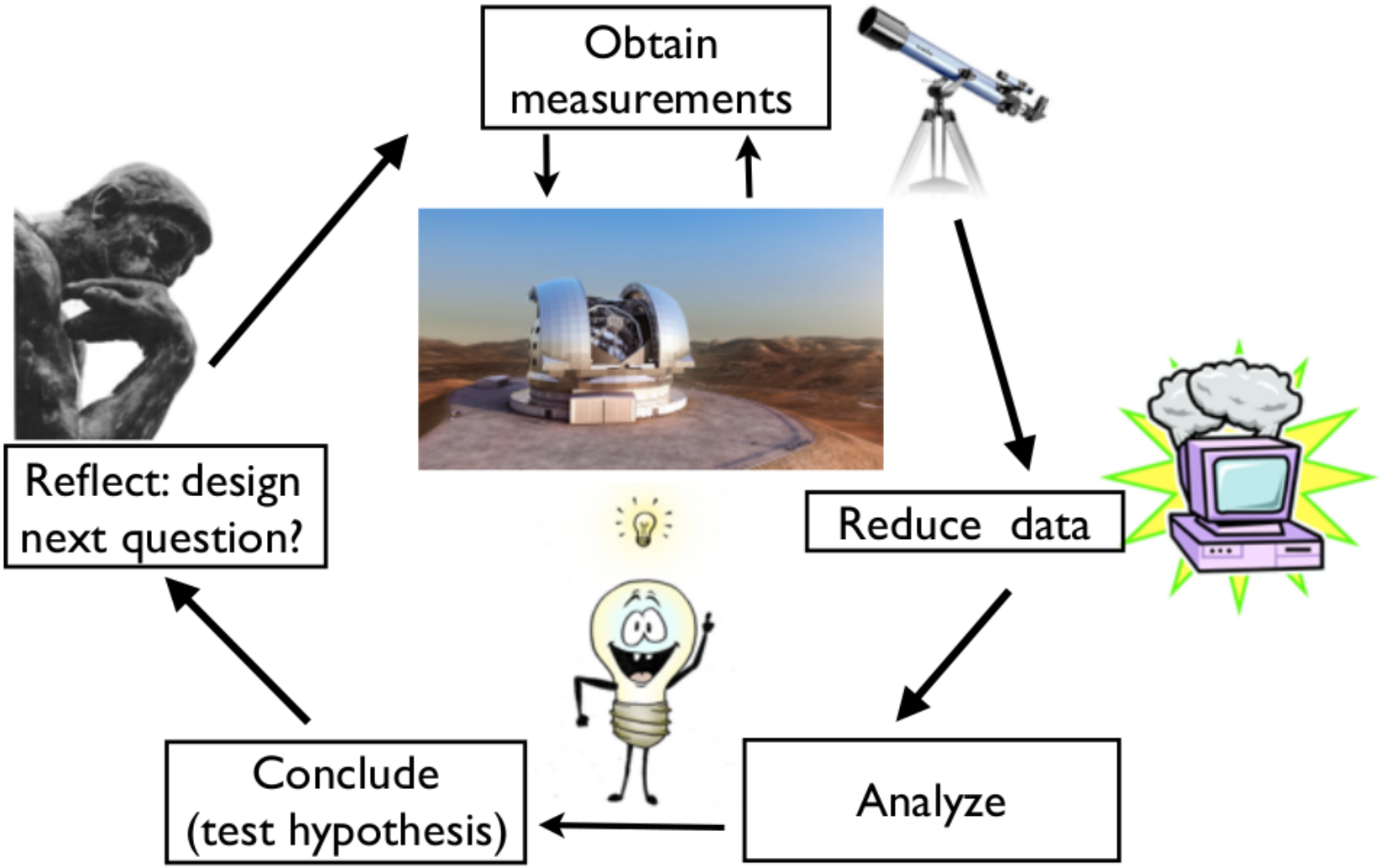
- Uma estatística é uma quantidade que sumariza os dados.
- É a redução de dados em última instância
- Pode ser um número (e.g. uma média) mas não necessariamente.
- É a base para se tomar uma decisão.

- **Qual é a particularidade da astronomia dentro da estatística?**
- A Astronomia é uma ciência observacional e não experimental e em várias circunstâncias o número de objetos de uma dada amostra é limitado (e.g. cosmologia). Nem sempre é possível "rolar um dado" mais vezes.

- Por conta da nossa impossibilidade de "refazer experimentos" e dado o tamanho geralmente pequenos das amostras, isto significa que não conhecemos necessariamente a distribuição subjacente das variáveis medidas.

- Ao não conhecer essa distribuição subjacente é difícil dizer se um resultado em particular é normal ou não.

# The process of science




8

# Como a ciência é feita?

Em termos simples cada experimento percorre um ciclo caracterizado por seis estágios:

1. **Observação:** dados são coletados via observação ou consulta a fontes já existentes (dados de arquivo, catálogos pré-existentes etc.)
2. **Processamento:** remoção de sinais instrumentais nos dados, como correção por *flat-field*, calibração fotométrica etc. Também é frequentemente chamado de redução de dados.
3. **Análise:** obtenção de números a partir dos dados livres de sinais instrumentais, como intensidades, posições, formas etc. Produção de descrições que sumarizem os dados e permitam comparações – estatísticas.
4. **Conclusão:** processo para se tomar uma decisão como teste de hipótese, correlação, modelamento etc.
5. **Interpretação:** o que foi aprendido? Essa conclusão/resultado é plausível? Se não o que deve ser re-checado? Se o resultado é inesperado, o que é preciso para confirma-lo?
6. **Desenho de novo experimento:** Se a questão é importante o suficiente e houverem condições novas, então um novo experimento poderá/deverá ser desenhado e proposto.

Table 1.1 *Stages in astronomy experimentation*

| Stage | How | Examples | Considerations |
|---|---|---|---|
| Observe | In person?<br>Remotely?<br>Depends on facility | Experiment design: calibration integration time<br>*Stats* | What is wanted?<br>Number of objects<br>*Stats* |
| Reduce | Algorithms | Flat field<br>Flux calibration | Data integrity<br>Signal-to-noise<br>*T Stats* |
| Analyse | Parameter estimation, hypothesis testing<br>*T Stats* | Intensity measurements<br>Positions<br>*T Stats* | Frequentist, Bayesian?<br>*T Stats* |
| Conclude | Hypothesis testing<br>*T Stats* | Correlation tests<br>Distribution tests<br>*T Stats* | Believable, repeatable, understandable?<br>*T Stats* |
| Reflect | Carefully; far too little time is invested here | Mission achieved? A better way? 'We need more data'?<br>*T Stats* | The next observations<br>*T Stats* |
| Design | Hone the mission; build science case<br>*Stats* | New observations/ instrument/ telescope/space mission | Feasibility – cost, team design, experience, human resources; simulations, predictions<br>*Stats* |

*Stats*: O uso de estatística ou inferência estatística é importante.

**T**: Importante para teóricos bem como para experimentais/observacionais.

# Probabilidades e distribuições probabilísticas

- Estatística e probabilidade são conceitos fortemente relacionados (mas não idênticos).
- Num mundo onde nossa estatística é derivada de quantidades finitas de dados a probabilidade é necessária para a inferência (tomada de decisão).
- Esses conceitos são naturais. Quando dizemos que alguém é alto, estamos levando em consideração a distribuição de altura de uma população e fazendo uma inferência que um determinado indivíduo é maior do que, por exemplo, 75% das pessoas da amostra.
- Nesse caso, a distribuição de alturas (normalizada para 1) é uma *função de densidade de probabilidade* ou *distribuição probabilística*.
- Existem formas matemáticas para essas distribuições como *Normal* ou *de Poisson* que veremos mais adiante.

# Tudo já resolvido?

- Existe uma tendência a se pensar que a estatística já está toda "resolvida", mas isso pode ser muito equivocado.
- Muitos fenômenos inesperados podem ocorrer e requerer o uso de estatísticas novas.
- A capacidade de perceber esses efeitos e formular novas estatísticas podem fazer toda a diferença num trabalho científico.

# Probabilidade e estatística em inferência

- Estatísticas são combinações dos dados que não dependem de parâmetros desconhecidos.
- A média é o exemplo mais comum.
- Quando calculamos a média de um conjunto de dados esperamos que isso tenha alguma relação com a média da distribuição subjacente a partir da qual os dados foram tomados.
- Classicamente calcula-se a distribuição amostral da média ou a probabilidade dos vários valores da média medida caso, hipoteticamente, repitamos o mesmo experimento várias vezes.
- Sabemos então qual a probabilidade que um dado intervalo ao redor de uma medida única vai conter a média verdadeira.

# O que é probabilidade

- Uma forma simples de pensar em probabilidade, (termo frequentemente associado a jogos de "azar" envolvendo cartas ou dados) é a seguinte:
- **Número de eventos favoráveis / Número total de eventos**
- Nesse contexto a probabilidade de obtermos um 6 com um dado é "obviamente" 1/6.
- Esse valor pode ser obtido supondo que a probabilidade do dado expor cada uma de suas faces é a mesma e tem valor $x$. Nesse caso tem-se que:

$$6x = 1 \rightarrow x = 1/6$$

- Esse cálculo aparentemente trivial se baseia em várias premissas:
  1. Princípio da Indiferença: a probabilidade de cada evento é a mesma, a menos de informações em contrário.
  2. Todos os resultados possíveis foram identificados. No caso os 6 lados do dado.
  3. Por convenção a soma de todas as probabilidades é a unidade.

# O que é probabilidade

- Embora o princípio da indiferença seja útil em diversas ocasiões não é sempre o caso.
- Vamos considerar, por exemplo, a probabilidade de que durante uma dada noite (quando um de vocês estiver tentando observar num telescópio) o céu estará nublado.
- Uma forma de estimar esse valor é a seguinte:
  **Número de noites nubladas no ano passado / 365**.
- Claramente duas complicações emergem
  1. A limitação dos dados. Seguramente 10 anos anteriores seriam melhores do que 1 apenas.
  2. A identificação de casos "igualmente prováveis". Para todas as noites a probabilidade do céu estar nublado seria igual. No LNA, por exemplo, é muito mais provável que as noites estejam "fechadas" no verão do que no inverno.
- O que é o "verão" nesse caso? Um conjunto de noites com igual probabilidade de estarem nubladas?

- Probabilidades só podem ser corretamente estimadas uma vez que foram identificados os casos igualmente prováveis e essa identificação é um passo subjetivo.

- É muito comum definir probabilidades a partir de afirmações empíricas baseadas em um grande número de casos (os 10 anos no exemplo acima).

- Mas como se viu, esse definição deve ser circular pois a seleção dos dados depende do conhecimento de quais casos são igualmente prováveis.

- Essa definição de probabilidade é chamada de frequentista.

- O que é então probabilidade? A notação que adotaremos nesse curso é uma formalização numérica de seu grau de intensidade em qualquer crença.

- Independentemente dessa definição subjetiva, é crítico que duas pessoas com as mesmas informações chegarão às mesmas probabilidades.

- Essa requerimento, expressado de modo apropriado, é suficiente para o desenvolvimento de uma teoria da probabilidade que é matematicamente equivalente àquela interpretada em termos frequentistas. Isso foi feito originalmente por Cox (1946, Am. J. Phys, 14, 1) e veremos a seguir.

- $A$, $B$ e $C$ são três eventos sobre os quais gostaríamos de medir o quão fortemente achamos que cada um pode acontecer. Nesse caso deve aplicar que: *se $A$ é mais provável que $B$ e $B$ é mais provável do que $C$, então $A$ é mais provável do que $C$.*

- Incrivelmente isso é suficiente para deduzir os axiomas da probabilidade propostos por Kolmogorov anos antes do artigo de Cox:

  1. Qualquer evento aleatório tem uma probabilidade $prob(A)$ entre 0 e 1;
  2. Um evento certo tem $prob(A) = 1$;
  3. Se $A$ e $B$ são eventos mutuamente exclusivos, então $prob(A \text{ ou } B) = prob(A) + prob(B)$.

- Esses axiomas são suficientes para todo o desenvolvimento da teoria matemática de probabilidade.

# Condicionalidade e independência

- Dois eventos são ditos *independentes* se a probabilidade de um não é afetada pelo que possamos saber a respeito da outra. Nesse caso, segue (de modo não trivial) dos axiomas do Kolmogorov que:

$$prob(A \text{ e } B) = prob(A) \times prob(B) \tag{1}$$

- No caso em que a independência não ocorre o que se quer saber é a *probabilidade condicional*: a probabilidade de $A$ dado que conhecemos $B$. A definição disso é:

$$prob(A|B) = \frac{prob(A \text{ e } B)}{prob(B)} \tag{2}$$

- Se $A$ e $B$ forem independentes a informação sobre se $B$ ocorreu ou não não deve afetar nossas crenças sobre $A$. Assim sendo $prob(A|B) = prob(A)$ e temos novamente que $prob(A \text{ e } B) = prob(A)\ prob(B)$

- Suponha que existam várias probabilidades para o evento $B$, então temos que:

$$prob(A) = \sum_i prob(A|B_i)prob(B_i) \tag{3}$$

- $A$ pode ser o parâmetro do modelo que se quer encontrar e os $B's$, por exemplo, parâmetros instrumentais que não são de interesse.

- Conhecendo as probabilidades $prob(B_i)$ podemos nos livrar desses *parâmetros irrelevantes* ou *nuisance parameters* através de uma soma ou integração.

- Esse processo é conhecido como *marginalização*.