

和沈老师讨论的进一步结果

数 33 赵丰

January 15, 2017

1 老师建议

检验方法要在已知概率模型的 01 序列上做，一方面通过仿真比较不同检验方法的误码率，另一方法要做适当的理论分析来说明哪种方法是误码率意义下更优的。然而实际处理数据只能在给定的 8 组数据的基础上，给生物系的张老师一个结果。我们能做的工作就是找一个合适的概率模型，在其基础上仿真和理论分析，寻找并说明某种检验方法在该概率模型下是更优的。为保证模型能在一定程度上反映原问题，一方面要建立从 01 序列到 RT value 的中间环节，另一方面要让概率模型产生的 01 序列与一般 RNA 的二级结构中的 01 序列有一定的相似性。为保证检验方法能在单位点数据量较小的情形下给出更好的判决，一方面我们要用上相邻位点的数据，另一方面我们要重复检验。

2 两条已知 RNA 的统计信息

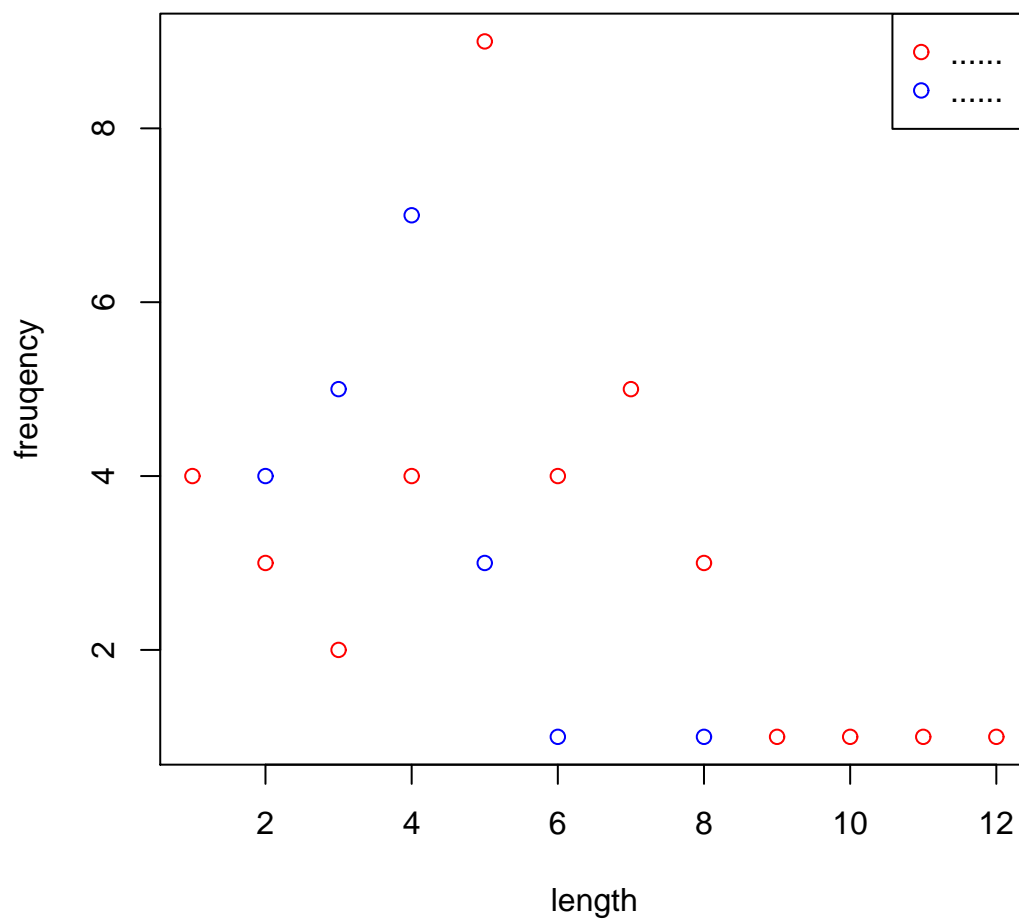
李盼给的第一个已知结构的 RNA 统计信息：

3 贝叶斯统计的方法

使用贝叶斯统计来做假设检验，需要有关于 01 序列的先验信息。假设 θ 服从 $p = \frac{1}{2}$ 的 Bernoulli 分布，总体 X 的条件分布 $(X|\theta = 0) \sim N(\mu_1, \sigma_1^2)$, $(X|\theta = 1) \sim N(\mu_2, \sigma_2^2)$, 现假设从总体中抽取了 n 个样本 X_1, X_2, \dots, X_n , X_i 是独立同分布的，根据以上信息做如下假设检验: $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ 欲使两类错误率总和最小，则拒绝域为 (具体推导见附录 1。):

$$\log(\sigma_1) + \frac{n(\bar{x} - \mu_1)^2}{\sigma_1^2} > \log(\sigma_2) + \frac{n(\bar{x} - \mu_2)^2}{\sigma_2^2} \quad (1)$$

.....RNA,.....299



根据贝叶斯统计的方法来做假设检验需要有 μ_1, μ_2 的先验信息，由于样本方差与样本均值的统计独立性，在缺少对方差 σ_1, σ_2 的先验信息的情况下，可以用样本方差对 σ_1, σ_2 做近似。进一步的，假设 $\sigma_1 = \sigma_2 = \sigma, \mu_2 > \mu_1$ ，可以推导出用贝叶斯统计的方法犯两类错误的概率相等，均为 $P(Z > \frac{\mu_2 - \mu_1}{2\sigma/\sqrt{n}})$ (具体推导见附录 2。)

4 Appendix 1: 拒绝域表达式的推导

由贝叶斯公式：

$$P(\theta = 0|\bar{x}) = \frac{p(\bar{x}|\theta = 0)P(\theta = 0)}{p(\bar{x}|\theta = 0)P(\theta = 0) + p(\bar{x}|\theta = 1)P(\theta = 1)} \quad (2)$$

$$= \frac{\frac{1}{\sqrt{2\pi\sigma_1^2/n}} \exp(-\frac{(\bar{x}-\mu_1)^2}{\sigma_1^2/n})}{\frac{1}{\sqrt{2\pi\sigma_2^2/n}} \exp(-\frac{(\bar{x}-\mu_2)^2}{\sigma_2^2/n}) + \frac{1}{\sqrt{2\pi\sigma_2^2/n}} \exp(-\frac{(\bar{x}-\mu_2)^2}{\sigma_2^2/n})} \quad (3)$$

同理可求出

$$P(\theta = 1|\bar{x}) = \frac{\frac{1}{\sqrt{2\pi\sigma_2^2/n}} \exp(-\frac{(\bar{x}-\mu_2)^2}{\sigma_2^2/n})}{\frac{1}{\sqrt{2\pi\sigma_2^2/n}} \exp(-\frac{(\bar{x}-\mu_2)^2}{\sigma_2^2/n}) + \frac{1}{\sqrt{2\pi\sigma_2^2/n}} \exp(-\frac{(\bar{x}-\mu_2)^2}{\sigma_2^2/n})} \quad (4)$$

根据观测数据当 $P(\theta = 0|\bar{x}) < P(\theta = 1|\bar{x})$ 时拒绝原假设 H_0 ，两式分母相同，对分子取对数即得到拒绝域的表达式。

5 Appendix 2: 贝叶斯统计方法犯两类错误的概率

设 R 为 (1) 式定义的拒绝域，则犯第一类错误的概率：

$$P(x \in R, \theta = 0) = P((\bar{X} - \mu_1)^2 > (\bar{X} - \mu_2)^2 | \bar{X} \sim N(\mu_1, \sigma^2)) \quad (5)$$

$$= P(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_2 - \mu_1}{2\sigma/\sqrt{n}} | \bar{X} \sim N(\mu_1, \sigma^2)) \quad (6)$$

同理，犯第二类错误的概率：

$$P(x \notin R, \theta = 1) = P(\frac{\bar{X} - \mu_2}{\sigma/\sqrt{n}} < \frac{\mu_1 - \mu_2}{2\sigma/\sqrt{n}} | \bar{X} \sim N(\mu_2, \sigma^2)) \quad (7)$$

因为 $\frac{\bar{X}-\mu_1}{\sigma}$ 服从标准正态分布 $Z \sim N(0, 1)$ ， Z 的分布具有对称性，由假设 $\mu_1 < \mu_2$ ，注意到 $\frac{\mu_2 - \mu_1}{2\sigma} > 0, \frac{\mu_1 - \mu_2}{2\sigma} < 0$ 所以贝叶斯统计的方法在方差相等的假设下犯两类错误的概率相等。

6 参考文献

References

- [1] 生物信息学