

Stat, fast

(I already said that)

July 31, 2016

0 Synopsis

Come ready to be put on the spot. You'll be less put on the spot if you know the following.

1 Probs

- [Permut/Combin]ation
- Choose
- Random Variables
- Discrete Distributions
- Continuous Distributions
- Joint Distributions
- $E[aX]$, $\text{Var}[X]$, $E[XY + Z]$ and $\text{Var}[aX + bY]$
- Covariance Vs Correlation Vs Causation
Vs Independent Vs Mutually Exclusive
- Bayes' Theorem
- Standard Error \downarrow

2 Inf(erence)

- MLE
- CLT
- Bootstrapping Vs Bayes
- Vs Confidence Intervals Vs Nonparametric (NP) estimation Vs NP tests \downarrow

3 Testing

- \emptyset Vs H_a
- p Vs α Vs Type II
- Multiple testing correction via Bonferroni
[Vs False Discovery Rate (FDR)]
- (A/B) t -test
- χ^2 -test
- $X + Y \sim ?$
- Bayes \downarrow

4 Multi-armed Bandit

- Ready, go

5 Bayes

We are monitoring credit card (cc) purchases for fraud by testing a measure of “unusual purchases”.

Let π denote the overall fraud rate (for which we may have some prior belief),

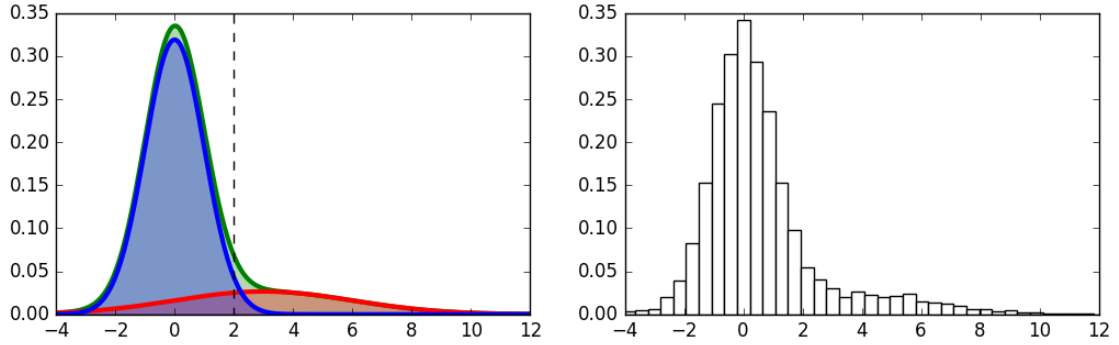
$$\begin{cases} f_i = 1 : & \text{if there is in fact fraud for cc } i, \text{ and} \\ f_i = 0 : & \text{if not,} \end{cases}$$

and t_i be our “unusual purchases” measure which will depend on f_i . That is,

$$\begin{aligned} d(\pi) &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1} \\ p(f_i|\pi) &= \pi^{f_i}(1-\pi)^{(1-f_i)} \\ t_i|f_i &\sim d_{f_i}(t_i) \end{aligned}$$

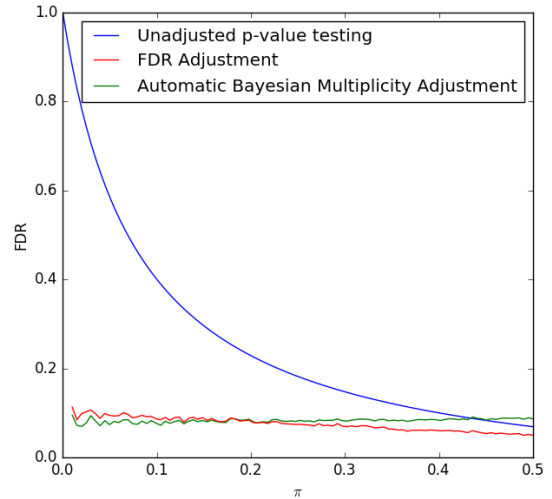
so that (upon marginalizing out the latent fraud indicator f_i) the measures t_i are generated from mixture distribution of fraudulent d_1 and genuine d_0 cc purchases

$$d(t_i|\pi) = (1-\pi)d_0(t_i) + \pi d_1(t_i)$$



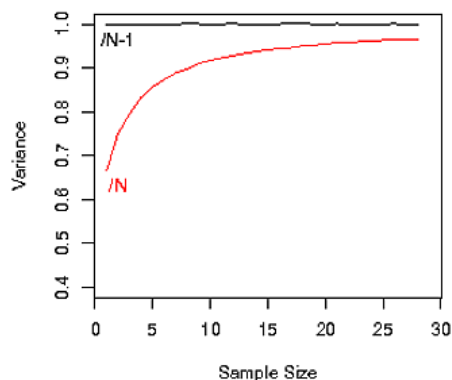
and we are interested in

$$\begin{aligned} &p(f_i = 1|t_1, \dots, t_n, f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n) \\ &\propto \prod d_{f_i}(t_i) p(f_i|\pi) \cdot d(\pi) \\ &\propto d_1(t_i) \int \pi^{1+\sum_{j \neq i} f_j + \alpha - 1} (1-\pi)^{\sum_{j \neq i} (1-f_j) + \beta - 1} d\pi \\ &\approx d_1(t_i) \hat{\pi} \\ &p(f_i = 0|t_1, \dots, t_n, f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n) \\ &\propto \prod d_{f_i}(t_i) p(f_i|\pi) \cdot d(\pi) \\ &\propto d_0(t_i) \int \pi^{\sum_{j \neq i} f_j + \alpha - 1} (1-\pi)^{1+\sum_{j \neq i} (1-f_j) + \beta - 1} d\pi \\ &\approx d_0(t_i) (1 - \hat{\pi}) \end{aligned}$$



False Discovery Rate (FDR), see: `statsmodels.sandbox.stats.multicomp.fdr correction0`

Appendix: n-1?



Dividing by $(n - 1)$ rather than n results in an unbiased estimator of σ^2

$$\begin{aligned}
 & E \left[\sum_{i=1}^n \left(x_i^2 - \frac{1}{n} \sum_{j=1}^n x_j \right)^2 \right] \\
 &= E \left[\sum_{i=1}^n \left(x_i^2 - \frac{2x_i}{n} \sum_{j=1}^n x_j + \left(\frac{1}{n} \sum_{j=1}^n x_j \right)^2 \right) \right] \\
 &= E \left[\sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n \sum_{j=1}^n x_i x_j + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n x_i x_j \right] \\
 &= E \left[\sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{j \neq i} x_i x_j + \frac{1}{n} \sum_{i=1}^n x_i^2 + \frac{1}{n} \sum_{j \neq i} x_i x_j \right]
 \end{aligned}$$

$$\begin{aligned}
 &= E \left[\sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{i=1}^n x_i^2 + \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{2}{n} \sum_{j \neq i} x_i x_j + \frac{1}{n} \sum_{j \neq i} x_i x_j \right] = E \left[\frac{n-1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{j \neq i} x_i x_j \right] \\
 &= \frac{n-1}{n} \sum_{i=1}^n E[x_i^2] - \frac{1}{n} \sum_{j \neq i} E[x_i x_j] \\
 &= \frac{n-1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \frac{1}{n} \sum_{j \neq i} \mu^2 \quad (\text{why?}) \\
 &= (n-1)(\sigma^2 + \mu^2) - \frac{n^2 - n}{n} \mu^2 = (n-1)\sigma^2
 \end{aligned}$$

Appendix: uncorrelated $\begin{matrix} \Rightarrow \\ ? \\ \Leftarrow \end{matrix}$ independent *hint*

