Notes on *Probability and Statistics (4th Edition)*

April 7, 2016

Bonferroni Inequality

$$
\begin{aligned}
P(\bigcup_{i=1}^{n} A_i) &\le \sum_{i=1}^{n} P(A_i) \\
P(\bigcap_{i=1}^{n} A_i) &\ge 1 - \sum_{i=1}^{n} P(A_i^c)
\end{aligned}
\tag{1}
$$

CDF

Properties:

- Nondecreasing
- $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$
- Continuity from right: $F(x) = (x^+)$ for all x

Probability Integral Transformation

Let X have **continuous** CDF $F(x)$, let $Y = F(X)$ (this is the transformation), then Y has the uniform distribution on $[0, 1]$.

(This can be proved by the properties of CDF.)

Conversely, if Y has the uniform distribution on $[0, 1]$, and F is a **continuous** CDF with quantile function F^{-1}, then $X = F^{-1}(Y)$ has the distribution with CDF F.

We can use this transformation to generate samples from a desired distribution with the help of a uniform distribution, e.g. generate samples from normal distribution with `random.uniform` function.

Distribution of Functions of Random Variables

Single Variable

Let X with PDF f and $P(a < X < b) = 1$ (a, b can be finite or infinite), let $Y = r(X)$, where r is **differentiable** and **one-to-one** on (a, b), let (α, β) be the image of (a, b) under r, let $s = r^{-1}$, then the PDF of Y is

$$g(y) = \begin{cases} f[s(y)]|\frac{ds(y)}{dy}| & \text{for } \alpha < y < \beta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Linear Function of Two Variables

Let X_1, X_2 with joint PDF $f(x_1, x_2)$, let $Y = a_1 X_1 + a_2 X_2 + b$ where $a_1 \neq 0$, then Y has a continuous distribution with PDF

$$g(y) = \int_{-\infty}^{\infty} f\left(\frac{y - b - a_2 x_2}{a_1}, x_2\right) \frac{1}{|a_1|} dx_2 \tag{3}$$

Transformation of Multiple Variables

(Transform n variables simultaneously.)

Let $X_1, \ldots, X_n$ have a continuous joint PDF $f(x_1, \ldots, x_n)$ where $P((X_1, \ldots, X_n) \in S) = 1$.

Let $r_1, \ldots, r_n$ be differentiable and one-to-one functions from S to T, which transform $X_1, \ldots, X_n$ to:

$$\begin{aligned} Y_1 &= r_1(X_1, \ldots, X_n) \\ Y_2 &= r_2(X_2, \ldots, X_n) \\ &\vdots \\ Y_n &= r_n(X_n, \ldots, X_n) \end{aligned} \tag{4}$$

Let $s_1, \ldots, s_n$ be invserses, which transform them back:

$$\begin{aligned} x_1 &= s_1(y_1, \ldots, y_n) \\ x_2 &= s_2(y_2, \ldots, y_n) \\ &\vdots \\ x_n &= s_n(y_n, \ldots, y_n) \end{aligned} \tag{5}$$

Then the joint PDF of transformed variables is:

$$g(y_1, \ldots, y_n) = \begin{cases} f(s_1, \ldots, s_n)|J| & \text{for } (y_1, \ldots, y_n) \in T \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where J is the determinant of Jacobian:

$$J = det \begin{bmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{bmatrix} \tag{7}$$

Linear Transformation of Multiple Variables

Let $\mathbf{X} = (X_1, \ldots, X_n)$ (random vector) has continuous joint PDF f, define $\mathbf{Y} = \mathbf{AX}$, then $\mathbf{Y}$ has a continuous joint PDF

$$g(\mathbf{y}) = \frac{1}{|det\mathbf{A}|} f(\mathbf{A}^{-1}\mathbf{y}) \tag{8}$$

Jensen Inequality

Let f be a convex function, let X have finite mean, then $E(f(X)) \geq f(E(X))$

Skewness

Let X have mean μ and standard deviation σ, and finite 3rd moment, then the skewness of X is

$$\frac{E((X-\mu)^3)}{\sigma^3} \tag{9}$$

Skewness measures the **lack of symmetry**.

Mean and Median

Let X have finite variance σ^2, let $\mu = E(X)$, then for any d, $E((X-\mu)^2) \leq E((X-d)^2)$. Let m be the median of the distribution of X, then $E(|X-m|) \leq E(|X-d|)$.

That is, **mean** minimizes **mean squared error**, **median** minimizes **mean absolute error**.

Covariance and Correlation

Schwarz Inequality

For U, V such that $E(UV)$ exists,

$$E(UV)^2 \leq E(U)^2 E(V)^2 \tag{10}$$

If the RHS is finite, then the **equality holds if and only if** there exist nonzero constants a, b such that $P(aU + bV = 0) = 1$.

Cauchy-Schwarz Inequality

For X, Y with finite variance,

$$Cov(X, Y)^2 \leq \sigma_X^2 \sigma_Y^2 \tag{11}$$

and

$$-1 \leq \rho(X, Y) \leq 1 \tag{12}$$

The **equality holds if and only if** there exists nonzero constants a, b, and constants c such that $P(aX + bY = c) = 1$.

Correlation

Caveats:

- Independent $\rightarrow$ uncorrelated
- Correlation only measures **linear** relationship

Conditional Expectation

Total probability of expectation and variance:

$$\begin{aligned} E[E(Y|X)] &= E(Y) \\ Var(Y) &= E[Var(Y|X)] + Var[E(Y|X)] \end{aligned} \tag{13}$$

Special Distributions

- Hypergeometric
  - **sampling without replacement**: A red balls, B blue balls, draw n balls with x red ones
  - **relation with binomial**: if n is negligible compared to $A+B$, then hypergeometric distribution is close to binomial with parameter n and $p = \dfrac{A}{A+B}$
- Poisson
  - close to binomial if $\lim_{n\to\infty} f_{Bin}(x|n, p_n) = f_{Poisson}(x|\lambda)$
  - close to hypergeometric if $\lim_{T\to\infty} \dfrac{n_T A_T}{A_T + B_T} = \lambda$
  - **Poisson process** with rate λ
    * number of arrivals in every fixed time interval of length t has Poisson distribution with mean λt
    * number of arrivals in every collection of disjoint time intervals are independent
- Negative Binomial
  - **failures before a certain number of successes**: the number of failures before the r-th success in a series of Bernoulli experiments with success probability p
- Geometry
  - **failures before the 1st success**: negative binomial with $r = 1$
  - r geometrically distributed variables with parameter p add up to a negative binomial variable with parameters r, p
  - **memoryless**: $P(X = k + t | X \geq k) = P(X = t)$, i.e., we only care about the value from a certain staring point
- Normal
  - the linear combination of independent normal variables is normal
- Lognormal
  - X has lognormal distribution if $\ln X$ has normal distribution
  - PDF $f(x) = \dfrac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(\ln x - \mu)^2}{2\sigma^2}\right)$
- Gamma
  - gamma function
    * $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$
    * **recurrence relation**: if $\alpha > 1$ then $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
    * **relation to factorial**: for every $n > 0$, $\Gamma(n) = (n-1)!$

  - **Stirling's Formula**: $\lim_{x \to \infty} \frac{(2\pi)^{1/2} x^{x-1/2} e^{-x}}{\Gamma(x)} = 1$
  - k independent gamma variables with α_i, β add up to gamma distribution with $\alpha_1 + \cdots + \alpha_k, \beta$
- Exponential
  - gamma distribution with $\alpha = 1$
  - **memoryless**: let $t > 0$, then for every $h > 0$, $P(X \geq t + h | X \geq t) = P(X \geq h)$
  - **distribution of minimum**: if $X_1, \ldots, X_n$ are from exponential distribution with β, then $Y = min\{X_1, \ldots, X_n\}$ has exponential distribution with $n\beta$
  - sort n exponential variables ascendingly $Z_1, \ldots, Z_n$, then $Y_k = Z_k - Z_{k-1}$ has exponential distribution with $(n + 1 - k)\beta$ (memoryless)
  - **relatino to Poisson, time interval between Poisson arrivals**: suppose arrivals according to Poisson process with rate β, let Z_k be the **time until the k-th arrival**, define $Y_1 = Z_1, Y_k = Z_k - Z_{k-1}$, then all Y_is are i.i.d. and each has the exponential distribution with β
  - **gamma and interval between Poisson arrivals**: just add up the exponentially distributed variables in the last list item
- Beta
  - beta function
    * $\mathrm{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$
    * $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
  - **conditional probability and binomial**: if P $\mathrm{B}(\alpha, \beta)$, and $X|P = p$ $Bin(n, p)$, then $P|X = x$ $\mathrm{B}(\alpha + x, \beta + n - x)$
- Multinomial
  - n **occurences of** k **events, event** i **occurs** x_i **times**
  - just the generalization of binomial

Hypotheses Testing

Basics

Regions:

- **Critical region**: the range of a **sample** that tells us to reject H_0
- **Rejection region**: the range of a **statistic** that tells us to reject H_0

The **power function** $\pi(\theta|\delta)$ describes the **probability** that we reject H_0 as a function of the parameter θ, given the test procedure δ.

If S_1 is the critical region, then $\pi(\theta|\delta) = P(\mathbf{X} \in S_1|\theta)$.

Two types of erros:

- Type I error: we reject a true H_0, $\pi(\theta|\delta)$ for $\theta \in \Omega_0$
- Type II error: we do not reject a false H_0, $1 - \pi(\theta|\delta)$ for $\delta \in \Omega_1$

If $\pi(\theta|\delta) \leq \alpha_0$ for all $\theta \in \Omega_0$, then the **level of significance** of the test δ is α_0, and the **size** of the test $\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta)$.

In other words:

- Given a test, the level or size tells us the upper bound of the probability of making type I error.
- Why *significance*? If two sets of data are different significantly, we can think that they are not from the same population. That's just the reason we reject H_0, since the data we observed do not have the desired property.
- If the level of significance is low, then it is less probable we would make type I error, which indicates that it is high probable that the difference exists, which further indicates that we should reject H_0.

P-value is a god-damn fucking concept.

Recall the normal test procedure:

1. we are told to satisfy some level of significance α_0
2. based on α_0, we assume H_0 is true, and determine the rejection region in the distribution $p(T|H_0)$
3. once we have the observation, we check if it falls within the rejection region

Now, forget α_0, **if we have the observation, and assume H_0 is true**, we can compute **the probability that the observation is significantly different from H_0** based on the unknown level α. Because this value can tell us to what extent we will make type I error.

We do the following:

1. determine the rejection region based on α
2. to compute the probability of significant difference, we treat the observation as if it falls within the rejection region
3. we adjust α to see how small it can achieve

Then the minimum α is the fucking p-value.

If the p-value $\leq \alpha_0$, then it tells us that we are $(1 - \alpha)$-confident (more confident than the given α_0) about the conclusion that the observation is significantly different from the hypothesis, so that we should reject H_0.

Or another way to think:

α_0 说你这个检验犯 type I 错的概率不能超过这么多（检验不能太不靠谱了）。而我假定 H_0 是对的，却观察到一组数据，基于这数据，我可以推断说，我拒绝 H_0 犯 type I 错的概率（因为假设 H_0 是对的）最小也能低于或等于 α_0。那我肯定要拒绝 H_0 了啊。

Also, according to Wikipedia:

The p-value is defined as **the probability**, under the assumption of hypothesis H_0, **of obtaining a result equal to or more extreme than what was actually observed**.

The smaller the p-value, the larger the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation.

Likelihood Ratio Test is to use the statistic $\Lambda(\mathbf{x}) = \dfrac{\sup_{\theta \in \Omega_0} f_n(\mathbf{x}|\theta)}{\sup_{\theta \in \Omega} f_n(\mathbf{x}|\theta)}$ and to reject H_0 if $\Lambda(\mathbf{x}) \leq k$ for some constant k.