بررسی آمارگان رنها و بکارگیری آن در تعدادی تست آماری

محمد دخیل‌علیان
استادیار دانشکده برق و کامپیوتر
دانشگاه صنعتی اصفهان
mdalian@cc.iut.ac.ir

چکیده

آزمونهای آماری ابزار مناسبی برای بررسی خواص دنباله‌های شبه تصادفی خصوصاً در سیستم‌های رمزنگاری می‌باشد. در این راستا آزمونهای متنوعی ارایه شده‌اند که از جمله آنها می‌توان به آزمون رنها[1] اشاره نمود. در این مقاله ضمن اشاره به آزمون متداول برای بررسی رنهای موجود در یک دنباله، با بررسی تئوری آمارگان رنها در یک دنبالهٔ باینری تصادفی ایده‌آل، روشی برای آزمودن رنها و علاوه بر آن آزمونهایی برای گپها[2] و بلوکهای[3] ارایه شده است.

کلمات کلیدی :
آزمونهای آماری، آزمون رنها، دنباله‌های شبه تصادفی، سیستمهای رمزنگاری.

۱- مقدمه

در بسیاری از کاربردها، استفاده از دنباله‌های کاملاً تصادفی از اهمیت قابل توجهی برخوردار می‌باشد. منظور از یک دنبالهٔ تصادفی، دنباله‌ای با عناصر مستقل و دارای توزیع یکسان و یکنواخت (*i.i.d*) می‌باشد، لذا مولدهایی مفیدتر می‌باشند که از لحاظ آماری دنباله‌هایی را با ویژگی‌های دنباله‌های تصادفی تولید نمایند.

با شناخت خواص آماری دنباله‌های تصادفی، معیارهایی برای دنباله‌های شبه تصادفی بیان شده است *[1]* و متعاقب آن نیز آزمونهای مختلفی برای بررسی رفتار تصادفی دنباله‌ها و مولدهایشان طرح شده‌اند. آزمونهای آماری برروی بخش کوچکی از دنباله انجام می‌شود و طی آن، دنباله در آزمون قبول یا رد می‌گردد. برای ارزیابی یک مولد معمولاً تعداد زیادی دنباله مورد آزمایش قرار می‌گیرد و لازم است تعداد معنی داری از دنباله‌ها، از آزمون عبور نمایند. ذکر این نکته ضروری است که قبول یا رد شدن دنباله‌ای در یک آزمون دلیل بر تصادفی بودن یک دنباله نیست، زیرا در این آزمونها رفتار نوعی دنباله‌های تصادفی مد نظر قرار می‌گیرد و بدین جهت دنباله‌هایی که با این رفتار نوعی مطابقت داشته باشند، مورد قبول واقع می‌گردند. آزمون رنها [۲-۳] از جمله آزمونهایی است که در این مقاله مورد بررسی قرار می‌گیرد و پیرو آن آزمونهای دیگری در این خصوص ارایه می‌گردد.

---

[1] *Run.*
[2] *Gap.*
[3] *Block.*

۲- آزمون رنها

در دنباله‌های کاملاً تصادفی باینری بیتهای صفر و یک بدون نظم و شکل معینی رخ می‌دهند و بدین لحاظ مدلسازی این دنباله‌ها برای تحقق یک آزمون آماری ضروری است. یکی از مشخصه‌های یک دنباله باینری تعداد و طول رنهای موجود در یک دنباله می‌باشد. به تعدادی بیت صفر و یا بیت یک که پشت سرهم در دنباله قرار گیرند، یک رن گفته می‌شود. اگر رن شامل تعدادی یک باشد به آن بلوک و اگر رن شامل تعدادی صفر باشد به آن گپ گفته می‌شود. به عنوان نمونه دنبالهٔ *0100011000011111110* شامل هفت رن به طول‌های ۱ ، ۲ ، ۳ ، ٤ و ٦ می‌باشد که سه تای آن بلوک به طول‌های ۱ و ۲ و ٦ و چهارتای آن گپ و به طولهای ۱ و ۳ و٤ می‌باشد. در یک دنباله باینری کاملاً تصادفی نظیر $X^n=X_1,X_2,...,X_n$ ، تعداد رنها در حالت کلی دارای توزیع مشخصی می‌باشند. لذا میانگین و واریانس آنها به عنوان معیاری در آزمون ارایه شده در [۳] در نظر گرفته شده است. به عبارت دیگر ثابت شده است که در یک دنبالهٔ n بیتی (هنگامی که n به سمت بی‌نهایت میل کند) ، تعداد رنهای موجود در یک دنباله دارای توزیع نرمال با میانگین و واریانس زیر می‌باشد [٤]:

متوسط تعداد رنها $m = 1 + \frac{2n_0 n_1}{n}$ (۱)

واریانس تعداد رنها $\sigma^2 = \frac{(m-1)(m-2)}{n-1}$

که در رابطهٔ فوق n_0 ، n_1 و n به ترتیب تعداد صفرها، یکها و طول دنباله می‌باشد. در مرجع [۳] نیز به گونه‌ای دیگر رابطه (۱)، مورد استفاده قرار گرفته است. بدین ترتیب پارامتر آزمون مربع کای[1] برای بررسی توزیع آماری رنهای موجود در یک دنباله به صورت زیر ارایه شده است.

$$\chi^2_{run} = \left(\frac{N_R - m}{\sigma}\right)^2 \qquad (۲)$$

که در عبارت فوق N_R تعداد کل رنها در یک دنبالهٔ نوعی مورد آزمایش با طول n بیت می‌باشد. آزمون فوق بواسطهٔ اینکه تنها متوسط و واریانس تعداد کل رنها را مورد آزمون قرار می‌دهد، نمی‌تواند آزمون کاملی برای بررسی آماری تعداد رنها باشد. در بخش بعد ضمن بیان تئوری توزیع رنها در یک دنبالهٔ کاملاً تصادفی چگونگی اصلاح آزمون رنها را بیان می‌نماییم.

۳- محاسبهٔ تابع احتمال رنها

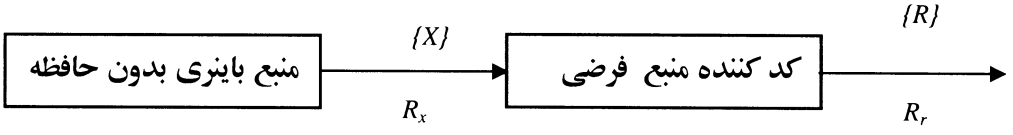
فرض کنید منبع S یک منبع باینری بدون حافظه باشد که بیتهای صفر و یک را با احتمال $\frac{1}{2}$ تولید کند، بنابراین آنتروپی منبع برابر یک بیت بر سمبل بوده و لذا سرعت اطلاعات این منبع برابر بیت ریت[2] منبع می‌باشد. حال اگر در خروجی این منبع از کد کنندهٔ منبع[3] فرضی شکل (۱) با مجموعه الفباء $r_2^0, r_1^1, r_1^0, ...$ به صورت جدول(۱) استفاده نماییم،کد کنندهٔ منبع دارای مجموعه الفبایی با بی‌نهایت عضو می‌باشد. توجه نمایید که علی‌رغم اینکه پیشوند بودن کلمات کد دراینجا مراعات نشده است، منتها فرض براین است که هیچگاه دو یا چند سمبل یکسان کنار هم کد

1

2 *Bit rate.*

3 *Source encoder.*

نشوند. به عنوان نمونه اگر دنبالهٔ ورودی برابر 011000 بود، دنبالهٔ خروجی تنها به $r_1^0 r_2^1 r_3^0$ کد گردد و هیچگاه دنباله خروجی به یکی از شکلهای $r_1^0 r_1^1 r_1^1 r_3^0$ یا $r_1^0 r_2^1 r_1^0 r_2^0$ و ... تبدیل نگردد.



شکل(۱)

جدول(۱): نگاشت خروجی منبع فرضی

سمبل‌های خروجی	ورودی
r_1^0	0
r_1^1	1
r_2^0	00
r_2^1	11
r_3^0	000
r_3^1	111
⋮	⋮

به این ترتیب سرعت اطلاعات در خروجی کد کنندهٔ منبع برابر R_x بیت بر ثانیه می‌گردد و با توجه به اینکه تعداد بیت در نظرگرفته برای هر سمبل برابر تعداد بیت ورودی است، لذا با توجه به اینکه بیتهای خروجی منبع باینری یکنواخت و مستقل می‌باشند ($H(X)=1$) ، لذا اگر سمبل‌های خروجی کد کنندهٔ منبع بخواهد وابسته به هم باشند، بالاجبار سرعت اطلاعات خروجی کمتر از سرعت اطلاعات ورودی می‌گردد ($R_r<R_x$) و این به معنی گم شدن اطلاعات یا ابهام در دکدینگ می‌باشد که با فرض بیان شده، این مسأله منتفی است، پس دنبالهٔ خروجی کد کنندهٔ منبع، دنباله‌ای با عناصر مستقل از هم می‌باشد. حال می‌خواهیم توزیع خروجی کد کنندهٔ منبع را محاسبه نماییم. فرض کنید $\{X\}$ دنبالهٔ تولید شده توسط منبع باینری باشد.

$$\{X\}=X_1, X_2, X_3, \ldots \qquad (۳)$$

برای تولید سمبل r_j^1 یا r_j^0 در خروجی کد کنندهٔ منبع، از ابتدای دنباله شروع می‌کنیم و لذا به سادگی می‌توان احتمال تولید سمبل‌های r_j^1 و r_j^0 را بدست آورد :

$$P(R=r_j^1)=P(X_i=1, X_{i+1}=1,\ldots, X_{i+j-1}=1, X_{i+j}=0) \qquad (۴)$$

با توجه به استقلال X_i ها از یکدیگر داریم :

$$P(R=r_j^1)=\left(\prod_{k=i}^{i+j-1} P(X_k=1)\right).P(X_{i+j}=0)=(\frac{1}{2})^{j+1} \qquad (۵)$$

به همین ترتیب می توان احتمال تولید سمبل r_j^0 را محاسبه نمود :

$$P(R = r_j^0) = P(X_i = 0, X_{i+1} = 0,..., X_{i+j-1} = 0\ , X_{i+j} = 1) = (\frac{1}{2})^{j+1} \qquad (۶)$$

لازم به ذکر است که در روابط فوق $X_{i+j}=0$ در نظر گرفته شده است، زیرا بنابر فرض هیچگاه دو یا چند سمبل یکسان در خروجی کد کننده مجاز نمی‌باشد و $X_{i+j}=0$ تضمین کنندهٔ این شرط می‌باشد. با توجه به روابط (۵) و (۶) می‌توان نتیجه گرفت، خروجی کد کنندهٔ منبع فرضی دنباله‌ای به صورت $\{R\}$ با سمبل‌های مستقل از هم می‌باشد و هر سمبل دارای تابع احتمال به صورت (۷) می‌باشد

$$\{R\} = R_1\ , R_2\ , R_3\ , \ldots \qquad (۷)$$

$$P(\ R_i = r_j^t\ ) = (\frac{1}{2})^{j+1}$$

$i=1, 2, \ldots \qquad t=0, 1$

واضح است که :

$$\sum_{\forall j,t} P(\ R_i = r_j^t\ ) = \sum_{t=0}^{1}\ \sum_{j=1}^{\infty}\ (\frac{1}{2})^{j+1} = 2\times(\frac{1}{4}+\frac{1}{8}+\ldots) = 1 \qquad (۸)$$

با این مدلسازی در واقع r_j^0 گپ به طول j و r_j^1 بلوک به طول j را در دنبالهٔ $\{X\}$ مشخص می‌کند. به عبارت دیگر رابطهٔ (۷) تابع احتمال گپها و بلوکها را در یک دنبالهٔ باینری کاملاً تصادفی مشخص کرده است. به سادگی می‌توان تابع احتمال رنها را نیز بدست آورد.

$$P(\ R = r_j\ ) = P(\ R = r_j^0\ \ or\ \ R = r_j^1\ ) = (\frac{1}{2})^{j+1} + (\frac{1}{2})^{j+1} = (\frac{1}{2})^{j} \qquad (۹)$$

به عبارت دیگر احتمال وقوع یک رن به طول j (R_j) برابر است با مجموع احتمال رخ دادن یک گپ به طول j یا بلوک به طول j. برای سادگی تابع احتمال برای رنها را به صورت زیر می‌نویسیم :

$P(R=R_j)=P(R=j)=P_R(j)$

$j=1,2,\ldots$

بنابر این متوسط طول هر رن را می‌توان به صورت زیر محاسبه نمود :

$$\overline{R} = \sum_{j=1}^{\infty} jP(\ R = j\ ) = \sum_{j=1}^{\infty} j(\frac{1}{2})^{j} = 2(\ bit\ ) \qquad (۱۰)$$

رابطهٔ فوق نشان می‌دهد که در یک دنبالهٔ باینری کاملاً تصادفی به طول n ، به طور متوسط $\frac{n}{2}$ رن وجود دارد که با توجه به روابط۵ الی ۷ ، به طور متوسط نصف رنها ($\frac{n}{4}$) گپ و نصف دیگر بلوک می‌باشد.

٤- چگونگی انجام آزمون رنها

برای انجام آزمون رنها روی یک دنبالهٔ باینری به طول n بیت نظیر X^n ، دنباله را به دنبالهٔ R^k تبدیل می‌کنیم.

$$X^n = X_1, X_2, \ldots, X_n \qquad (۱۱)$$

$$R^{k'} = R_1, R_2, \ldots, R_k$$

k با توجه به رابطهٔ۱۰ به طور متوسط باید حدود $\frac{n}{2}$ باشد. حال فرض کنید آزمون را برای رنهای به طول

حداکثر m بخواهیم انجام دهیم. برای اینکار رنهای به طول بزرگتر از m را به عنوان وقوع یک سمبل فرضی * در نظر می‌گیریم که احتمال وقوع آن با توجه به رابطهٔ (۹) برابر است با :

$$P(R_i = *) = P(R_i > m) = 1 - P(R_i \le m)$$
$$= 1 - (\frac{1}{2} + \frac{1}{4} + ... + \frac{1}{2^m}) = \frac{1}{2^m} \qquad (۱۲)$$

بنابر این در مجموع می‌توان گفت :

$$P(R_i = j) = \begin{cases} (\frac{1}{2})^j & j = 1,2,...,m \\ & i = 1,2,\cdots,k \\ \frac{1}{2^m} & j = * \end{cases} \qquad (۱۳)$$

بنابراین می‌توان گفت دنبالهٔ R^k یک متغییر تصادفی چند جمله‌ای با تابع احتمال رابطه (۱۳) می‌باشد. بنابراین طبق قضیه آزمون مربع کای[۵]، پارامتر آزمون رنها را می‌توان به صورت زیر نوشت :

$$\chi_r^2 = \sum_{i=1}^{m} \frac{(r_i - k.2^{-i})^2}{k.2^{-i}} + \frac{(r^* - k.2^{-m})^2}{k.2^{-m}} \qquad (۱۴)$$

درجهٔ آزادی متغیر تصادفی χ_r^2 برابر m می‌باشد و شرط انجام آزمون با توجه به تقریب توزیع مربع کای به صورت زیر می‌باشد :

$$\frac{n}{2} \times \frac{1}{2^m} > 5 \qquad \text{یا} \qquad n > 5 \times 2^{m+1} \qquad (۱۵)$$

یا به طور معادل اگر طول دنبالهٔ مورد آزمون برابر n باشد، مقدار m باید در شرط زیر صدق نماید :

$$m < \log_2(\frac{n}{10}) \qquad (۱۶)$$

(در رابطهٔ (۱۴)، r^* تعداد رنهای به طول بیش از m در دنبالهٔ R^k می‌باشد)

حال اگر آزمون به این صورت انجام شود که تنها رنهای تا طول حداکثر m مورد آزمون قرار گیرند و بخواهیم سمبل فرضی * را در نظر نگیریم :

$$P(R_i = j / j \le m) = \frac{P(R_i = j, j \le m)}{P(j \le m)} = \frac{P(R_i = j)}{\frac{1}{2} + \frac{1}{4} + ... + \frac{1}{2^m}} = \frac{2^{-j}}{1 - 2^{-m}} \qquad (۱۷)$$

به این ترتیب دنباله باینری X^n به دنباله R^L تبدیل می‌نماییم که تنها شامل رنهای به طول حداکثر m با تابع احتمال(۱۷) می‌باشد و واضح است که :

$$R^L = R_1, R_2, \cdots, R_L$$
$$\sum_{j=1}^{m} P(R_i = j / j \le m) = \sum_{j=1}^{m} \frac{2^i}{1 - 2^m} = 1 \qquad (۱۸)$$

به عبارت دیگر در دنبالهٔ R^L هر یک از R_i ها دارای توزیع زیر می باشند:

$$P(R_i = j) = \frac{2^{-j}}{1-2^{-m}} \qquad (۱۹)$$
$$j = 1,2,...,m$$

بنابراین پارامتر آزمون رنها را می‌توان در این وضعیت به صورت زیر نوشت :

$$\chi_r^2 = \sum_{i=1}^{m} \frac{(r_i - L\frac{2^{-i}}{1-2^{-m}})^2}{L\times\frac{2^{-i}}{1-2^{-m}}} \qquad (۲۰)$$

درجهٔ آزادی متغیر تصادفی χ_r^2 برابر $m\text{-}1$ می‌باشد. در حالت ایده‌آل به طور متوسط L نیز برابر است با :

$$E(L) = \frac{n}{2}\times\sum_{j=1}^{m} P(R_i = j) = \frac{n}{2}(\frac{1}{2}+\frac{1}{4}+...+\frac{1}{2^m}) = \frac{n}{2}(1-2^{-m}) \qquad (۲۱)$$

بدین ترتیب شرط انجام آزمون به صورت زیر بدست می‌آید که در واقع همان رابطهٔ (۱۵) می‌باشد. بدین ترتیب m باید در شرط رابطه (۱۶) نیزصدق نماید.

$$\frac{n}{2}(1-2^{-m})\times\frac{2^{-m}}{(1-2^{-5})} > 5 \Rightarrow n > 5\times 2^{m+1} \qquad (۲۲)$$

۵-آزمون‌های گپ و بلوک

اگر بخواهیم آزمونی برای گپها و آزمون دیگری به طور مجزا برای بلوکها ترتیب دهیم. به سادگی از روند ارایه شده در بخش ٤ می‌توان استفاده نمود: به این ترتیب که اگر دنبالهٔ n بیتی X^n توسط کد کنندهٔ منبع کد گردد و تنها گپ‌های موجود در خروجی مد نظر قرار گیرد (یا به طور معادل بلوکها) :

$$X^n = X_1, X_2, \ldots, X_n \qquad (۲۳)$$
$$G^k = G_1, G_2, \ldots, G_k$$

با توجه به نتیجهٔ رابطه (۱۰) ، مقدار k به طور متوسط باید حدود $\frac{n}{4}$ باشد. اگر فرض کنیم منظور از $P(G_i=j)$ یعنی احتمال اینکه گپ iام به طول j باشد، با توجه به حذف بلوکها در خروجی کد کنندهٔ منبع، تابع احتمال گپها به صورت زیر بدست می‌آید :

$$P(G_i = j) = \frac{1}{2^j} \qquad i = 1,2,\cdots,k \qquad j = 1,2,3,... \qquad (۲۴)$$

به طور مشابه برای بلوکها می‌توان رابطه‌ای مشابه با رابطهٔ (۲۴) نوشت :

$$B^{k'} = B_1, B_2, ..., B_{k'} \qquad (۲۷)$$
$$P(B_i = j) = \frac{1}{2^j} \qquad i = 1,2,\cdots,k' \qquad j = 1,2,...$$

همانطور که ملاحظه می‌گردد تابع احتمال بیان شده در (۲۴) و (۲۵) کاملاً مشابه رابطهٔ (۹) می‌باشد. بنابراین با روندی مشابه بخش ٤، می‌توان پارامترهای آزمون گپها و بلوکها را به صورتهای زیر بیان نمود:

$$\chi_g^2 = \sum_{i=1}^{m} \frac{(g_i - k.2^{-i})}{k.2^{-i}} + \frac{(g^* - k.2^{-m})^2}{k.2^{-m}} \qquad (۲۶)$$

$$\chi_b^2 = \sum_{i=1}^{m} \frac{(b_i - k'.2^{-i})}{k'.2^{-i}} + \frac{(b^* - k'.2^{-m})^2}{k'.2^{-m}} \qquad (۲۷)$$

در روابط فوق g_i و b_i به ترتیب تعداد گپها و بلوکهای به طول i و g^* و b^* مشخص کننده به ترتیب تعداد گپها و بلوکهایی که طول آنها بیش از m در دنباله‌های G^k و $B^{k'}$ می‌باشد. لازم به ذکر است که شرط انجام آزمونهای ارایه شده توسط روابط (۲۶) و (۲۷) عبارت است از :

$$k \times \frac{1}{2^m} > 5 \Rightarrow \frac{n}{4} \times \frac{1}{2^m} > 5 \Rightarrow n > 5 \times 2^{m+1} \qquad (۲۸)$$

یا مقدار m انتخابی باید در شرط زیر صدق کند

$$m < log_2(\frac{n}{20}) \qquad (۲۹)$$

شرط بیان شده در (۲۸) و (۲۹) برای آزمونهای گپ و بلوک به طور یکسان برقرار می‌باشد.

در صورتی که بخواهیم آزمون گپها و بلوکها برروی گپها و بلوکهای تا طول حداکثر m ، انجام شود و بخواهیم سمبل فرضی g^* و b^* را در نظر نگیریم، پارامترهای آزمون به صورت زیر تبدیل می‌گردند :

$$\begin{cases} G^L = G_1, G_2, ..., G_L \\ B^{L'} = B_1, B_2, ..., B_{L'} \end{cases} \qquad (۳۰)$$

$$\begin{cases} P(G_i = j) = \dfrac{2^{-j}}{1-2^{-m}} \\ i = 1,2,\cdots,L \\ P(B_i = j) = \dfrac{2^{-j}}{1-2^{-m}} \\ i = 1,2,\cdots,L' \end{cases} \qquad j = 1,2,...,m \qquad (۳۱)$$

$$\chi_g^2 = \sum_{i=1}^{m} \frac{(g_i - L\frac{2^{-i}}{1-2^{-m}})^2}{L \times \frac{2^{-i}}{1-2^{-m}}} \qquad \text{درجهٔ آزادی } m\text{-}1 \qquad (۳۲)$$

$$\chi_b^2 = \sum_{i=1}^{m} \frac{(b_i - L'\frac{2^{-i}}{1-2^{-m}})^2}{L' \times \frac{2^{-i}}{1-2^{-m}}} \qquad \text{درجهٔ آزادی } m\text{-}1 \qquad (۳۳)$$

شرط انجام آزمون و تقریب مناسب توزیع مربع کای برای χ_g^2 و χ_b^2 دقیقاً همان روابط (۲۸) و (۲۹) می‌باشد.

٦- نتیجه‌گیری

در یک آزمون آماری استفاده از تابع احتمال و ترتیب دادن آزمون براساس آن بسیار مناسب‌تر از آزمونی است که براساس متوسط و واریانس تابع احتمال طرح شده است. در این تحقیق به بررسی آمارگان و بیان توزیع رنها و به

تبع آن گپها و بلوکهای یک دنبالهٔ کاملاً تصادفی پرداخته شد و سپس براساس آن برای رنها، گپها و بلوکها دو روش آزمون ارایه گردید که یکی دارای درجه آزادی m و دیگری $m-1$ می‌باشد و بدین ترتیب می توان گپها و بلوکهای یک دنباله را بطور مجزا مورد آزمون قرارداد.

مراجع:

[1] S.W.Golomb,*Shift Register Sequences*,Holden-Day, San Fransisco,1982.

[2] Beker H. and Piper F.,*Cipher Systems: The Protection of Communication*, Northwood Book, London,1982.

[3] Rukhin,Soto and etc.,"*A Statistical Test suite for Random and Pseudorandon Generator for Cryptographic Application*",NIST Special Publication,2001.

[4] Kimberely M.," Comparison *of Two Statistical Tests of Keystream Sequences*",Electronic letter, Vol.23,No.8, PP.365-366, April 1987.

[5] Larson G.H., *Introduction to Probability and Statistical Inference*, John-Wiley, 1974.