پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

مدل سازی میانگین امتیازات یک کشتی‌گیر در مسابقات جهانی

نگارندگان:

محمد حسینی
سمیه آدینه
سارا معصومی مغانلو

استاد درس:

دکتر عبدالله صفری

پروژهٔ امتیازی درس احتمال۲

زمستان ۱۴۰۱

مرحله اول: داده‌های زیر $(x_1, x_2, \dots, x_{58})$ نشانگر امتیازات حسن یزدانی در مسابقات بین‌المللی در سال‌های ۲۰۱۵ تا ۲۰۲۱ هستند.

3	9	6	6	16	6	8	10	9	12	10	9	13	6	10
7	10	4	11	11	10	4	10	10	12	10	13	12	10	12
11	10	11	10	10	6	16	11	11	10	10	11	5	11	5
9	10	10	10	10	3	7	12	11	6	8	10	12		

فرض می‌کنیم $X_1, \dots, X_{58} \overset{iid}{\sim} N(\mu, \sigma^2)$ باشند.

۱. ابتدا می‌خواهیم پارامترهای توزیع نرمال یعنی μ, σ^2 را برآورد کنیم. برای برآورد کردن این دو پارامتر از روش $Maximum\ Likelihood\ Estimation$ استفاده می‌کنیم.

می‌دانیم:

$$\hat{\mu}_{Ml} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

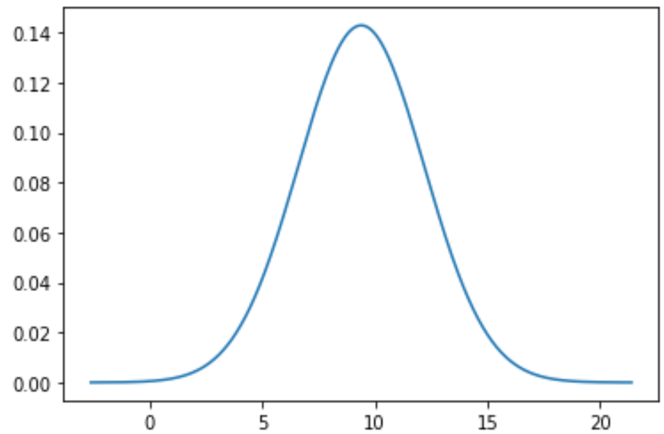
در نتیجه داریم:

$$\hat{\mu}_{Ml} = \frac{1}{58}\sum_{i=1}^{58} X_i = \frac{545}{58} = 9.4$$

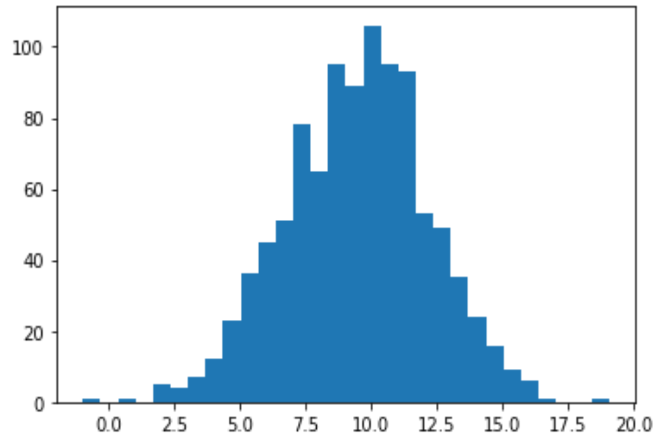
$$\hat{\sigma}^2_{ML} = \frac{1}{58}\sum_{i=1}^{58}(X_i - \bar{X})^2 = 7.8$$

با توجه به پارامترهای برآورد شده می‌توان گفت توزیع تقریبی $X_1, \dots, X_{58}$ به‌صورت $N(9.4, 7.8)$ است.

۲. در این بخش ۱۰۰۰ عدد تصادفی از توزیع $N(9.4, 7.8)$ تولید کردیم. شکل۱ نشانگر تابع چگالی احتمال این توزیع است و شکل۲ هیستوگرام ۱۰۰۰ عدد تصادفی تولید شده از این توزیع را نشان می‌دهد.

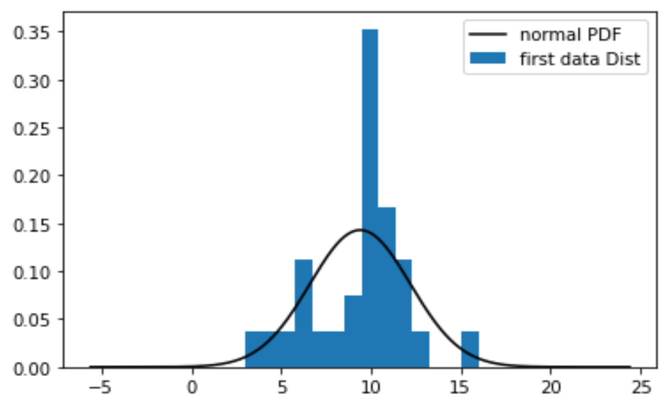


شکل ۱ – توزیع نرمال تخمین زده شده

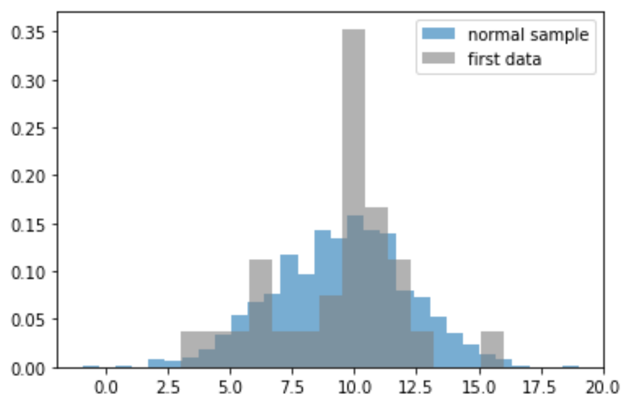


شکل ۲ – هیستوگرام داده‌های تولید شده

۳. حال می‌خواهیم توزیع داده‌های اولیه $(x_1, x_2, \dots, x_{58})$ را با توزیع تخمین زده شده و اعداد تصادفی تولید شده مقایسه کنیم.



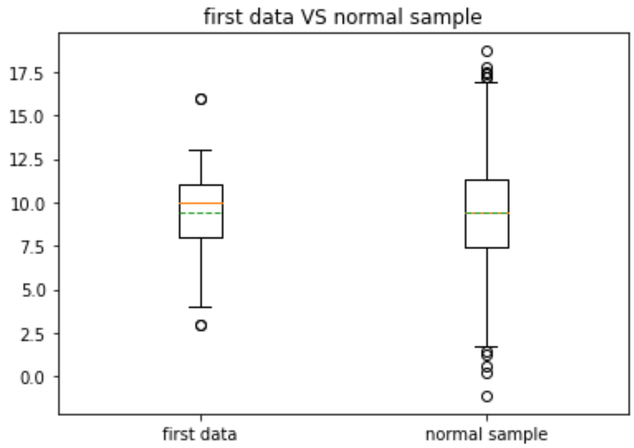
شکل ۳ – مقایسه توزیع نرمال و توزیع داده‌های اولیه



شکل ۴ – مقایسه توزیع داده‌های تولید شده و داده‌های اولیه

همانطور که از شکل۳و ۴ مشخص است، شکل ظاهری توزیع نرمال تخمین زده شده و توزیع داده‌های اولیه تفاوت‌های آشکاری دارد. برای مثال توزیع نرمال توزیعی متقارن است در حالی‌که توزیع داده‌های اولیه دارای تقارن نیست. البته مقدار میانگین و واریانس این دو توزیع نزدیک به هم است.

با توجه به نمودار جعبه‌ای می‌توان نتیجه گرفت که توزیع داده‌های اولیه نسبت به توزیع نرمال در بازهٔ کوچک‌تری متمرکز شده است. (یا به بیان دیگر می‌توان گفت توزیع نرمال دم‌های محتمل‌تری دارد.)



شکل ۵ – نمودار جعبه‌ای داده‌های تولید شده و داده‌های اولیه

مرحلهٔ دوم: در این مرحله به ۷ دادهٔ جدید از امتیازات این کشتی‌گیر در سال ۲۰۲۲ دست پیدا کرده‌ایم. این داده‌ها به‌صورت زیر هستند:

$$(y_1, y_2, \ldots, y_7) = (11, 12, 12, 1, 10, 11, 10)$$

همچنین می‌دانیم اگر متغیرهای تصادفی $Y_1, Y_2, \ldots, Y_7$ نشانگر مشاهدات جدید ما در سال ۲۰۲۲ باشند، خواهیم داشت:

$$Y_1, Y_2, \ldots, Y_7 | \theta, \lambda \overset{iid}{\sim} N(\theta, \lambda)$$
$$\theta \sim N(\mu, \sigma^2)$$
$$\lambda = 8.4$$

هدف اصلی ما این است که با استفاده از داده‌های موجود، میانگین امتیازات این کشتی‌گیر (θ) در مسابقات جهانی را مدل‌سازی کنیم. در مرحلهٔ قبل μ, σ^2 را برآورد کردیم؛ پس می‌دانیم توزیع تقریبی θ به‌صورت زیر است:

$$\theta \sim N(9.4, 7.8)$$

حال با استفاده از مشاهدات جدید می‌خواهیم باورهای خود نسبت به θ را به‌روز رسانی کنیم. در نتیجه علاقه‌مند به پیدا کردن توزیع $\theta | Y_1, Y_2, \ldots, Y_7$ یا همان $\theta | \underline{Y}$ هستیم.

۱. تابع چگالی احتمال $\theta | \underline{Y}$ به‌صورت زیر است:

$$f_{\theta|\underline{Y}}(t) = \frac{f_{\underline{Y}|\theta}(\underline{y}) f_\theta(t)}{\int_{-\infty}^{+\infty} f_{\underline{Y}|\theta=t}\left(\underline{y}\right) f_\theta(t) dt}$$

ابتدا صورت کسر را محاسبه می‌کنیم:

$$f_{\underline{Y}|\theta}\left(\underline{y}\right)=\prod_{i=1}^{7}\frac{1}{\sqrt{2\pi\lambda}}\ e^{-\frac{1}{2\lambda}(y_i-t)^2}$$ (به دلیل استقلال Y_i ها)

$$f_\theta(t)=\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$$

$$\rightarrow\ f_{\underline{Y}|\theta}\left(\underline{y}\right)f_\theta(t)=\ \frac{1}{\sqrt{2\pi\sigma^2}}\ \ e^{-\frac{1}{2\sigma^2}(t-\mu)^2}\times\prod_{i=1}^{7}\frac{1}{\sqrt{2\pi\lambda}}\ e^{-\frac{1}{2\lambda}(y_i-t)^2}$$

مخرج کسر نیز به صورت زیر به‌دست می‌آید:

$$\int_{-\infty}^{+\infty} f_{\underline{Y}|\theta=t}\left(\underline{y}\right)f_\theta(t)dt=\ \int_{-\infty}^{+\infty}\ \frac{1}{\sqrt{2\pi\sigma^2}}\ \ e^{-\frac{1}{2\sigma^2}(t-\mu)^2}\times\prod_{i=1}^{7}\frac{1}{\sqrt{2\pi\lambda}}\ e^{-\frac{1}{2\lambda}(y_i-t)^2}\ dt$$

$$=\int_{-\infty}^{+\infty}\ \frac{1}{\sqrt{2\pi\sigma^2}}\times\left(\frac{1}{\sqrt{2\pi\lambda}}\right)^7 exp\ (-\frac{1}{2\lambda}\sum_{i=1}^{7}(y_i-t)^2-\frac{1}{2\sigma^2}\,(t-\mu)^2)\ dt$$

$$\begin{cases}A=(\frac{1}{\sqrt{2\pi\lambda}})^7\times\frac{1}{\sqrt{2\pi\sigma^2}}\\ B=-\frac{1}{2\lambda}\\ C=-\frac{1}{2\sigma^2}\end{cases}$$

$$\rightarrow\int_{-\infty}^{+\infty} f_{\underline{Y}|\theta=t}\left(\underline{y}\right)f_\theta(t)dt=A\ \int_{-\infty}^{+\infty} exp\left\{B\sum_{i=1}^{7}(y_i-t)^2+C\ (t-\mu)^2\right\}dt$$

$$=A\ \int_{-\infty}^{+\infty} exp\left\{B(\sum_{i=1}^{7}y_i^2+7t^2-2t\sum y_i)+C(t^2+\mu^2-2\mu t)\right\}\ dt$$

$$\begin{cases} D = \sum_{i=1}^{7} y_i^2 \\ E = -2B \sum_{i=1}^{7} y_i \end{cases}$$

$$\rightarrow \int_{-\infty}^{+\infty} f_{\underline{Y}|\theta=t}\left(\underline{y}\right) f_\theta(t) dt = A \int_{-\infty}^{+\infty} exp\{(D + 7Bt^2 + Et) + C(t^2 + \mu^2 - 2\mu t\}\, dt$$

$$= A \int_{-\infty}^{+\infty} exp(C + 7B)\, t^2 + (E - 2C\mu)t + D + C\mu^2)\, dt$$

$$\begin{cases} F = C + 7B = -\frac{1}{z\sigma^2} + 7\left(-\frac{1}{z\lambda}\right) \\ G = E - 2C\mu = -2B \sum_{i=1}^{7} y_i - 2C\mu \\ H = D + C\mu^2 = -\frac{1}{2\lambda} \sum_{i=1}^{7} y_i^2 + (-\frac{1}{2z})\mu^2 \end{cases}$$

$$\mu = 9.3965,\ \ \sigma^2 = 7.7221,\ \ \lambda = 8.4000,\ \ \sum_{i=1}^{7} y_i^2 = 731,\ \ \sum_{i=1}^{7} y_i = 67$$

$$\rightarrow \begin{cases} F = -0.4814 \\ G = 9.1930 \\ H = -49.2289 \end{cases}$$

$$\rightarrow A \int_{-\infty}^{+\infty} exp(F\, t^2 + Gt + H)\, dt = A \times 2.6 \times 10^5$$

بنابرین:

$$f_{\theta|\underline{Y}}(t) = \frac{a\, .\, exp(Ft^2 + Gt + H)}{a \times 2.6 \times 10^5} = 0.3837 \times 10^{-5}.\, e^{Ft^2+Gt+H}$$

۲. اکنون که تابع چگالی احتمال $\theta|\underline{Y}$ را به‌دست آوردیم، در پی تخمین پارامترهای آن با توجه به پارامترهای برآورد شدهٔ بخش قبل هستیم. با توجه به فرم تابع می‌توان به نرمال بودن توزیع پی برد. بنابراین می‌خواهیم پارامترهای واریانس $(\acute{\sigma}^2)$ و میانگین $(\acute{\mu})$ را برآورد کنیم.

$$0.3837 \times 10^{-5}\, e^{Ft^2+Gt+H} =?\ \frac{1}{\sqrt{2\pi\acute{\sigma}^2}} e^{-\frac{1}{2\acute{\sigma}^2}(t-\acute{\mu})^2}$$

$$Ft^2 + Gt + H = -\frac{1}{2\acute{\sigma}^2}(t-\acute{\mu})^2 + N$$

$$\to Ft^2 + Gt + H = -\frac{1}{2\acute{\sigma}^2}t^2 + \frac{\acute{\mu}}{\acute{\sigma}^2}t - \frac{\acute{\mu}^2}{z\acute{\sigma}^2} + N$$

$$F = -\frac{1}{2\acute{\sigma}^2} \ \to\ \widehat{\acute{\sigma}^2} = -\frac{1}{2F} = 1.0386$$

$$G = \frac{\acute{\mu}}{\acute{\sigma}^2} \to \hat{\mu} = G\widehat{\acute{\sigma}^2} = 9.5479$$

$$H = -\frac{\hat{\mu}^2}{2\widehat{\acute{\sigma}^2}} + N \ \to\ N = H + \frac{\hat{\mu}^2}{2\widehat{\acute{\sigma}^2}} = 11.4840$$

$$\to\ f_{\theta|\underline{Y}}(t) = 0.3837 \times 10^{-5}\, exp\,(-\frac{1}{2\times1.04}(t-9.55)^2 + 11.4841)$$

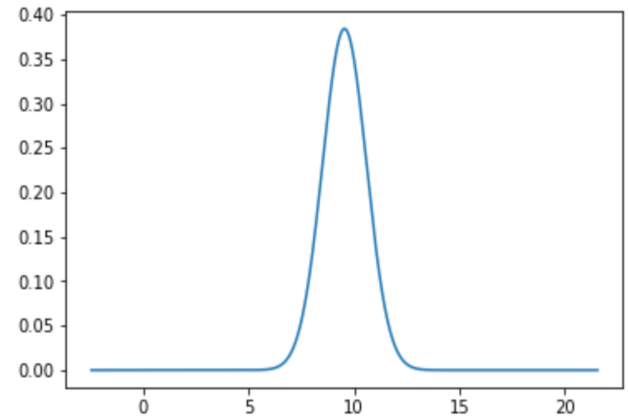
$$= 0.3915\, e^{-\frac{1}{2\times1.04}(t-9.55)^2} = \frac{1}{\sqrt{2\pi1.04}} e^{-\frac{1}{2\times1.04}(t-9.55)^2}$$

$$f_{\theta|\underline{Y}}(t) = \frac{1}{\sqrt{2\pi\times1.04}} e^{-\frac{1}{2\times1.04}(t-9.55)^2}$$

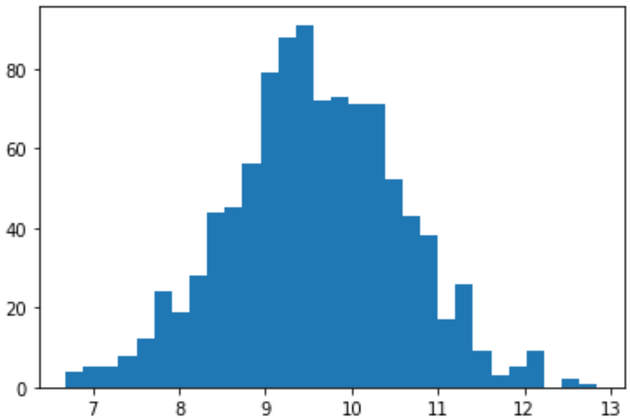
در نتیجه به‌دست آوردیم:

$$\theta|\underline{Y} \sim N(9.55, 1.04)$$

۳. در این مرحله از توزیع نرمال تخمین زده شده در بالا، ۱۰۰۰ عدد تصادفی تولید کردیم. شکل۶ نشانگر تابع چگالی احتمال این توزیع است و شکل۷ هیستوگرام ۱۰۰۰ عدد تصادفی تولید شده از این توزیع را نشان می‌دهد.

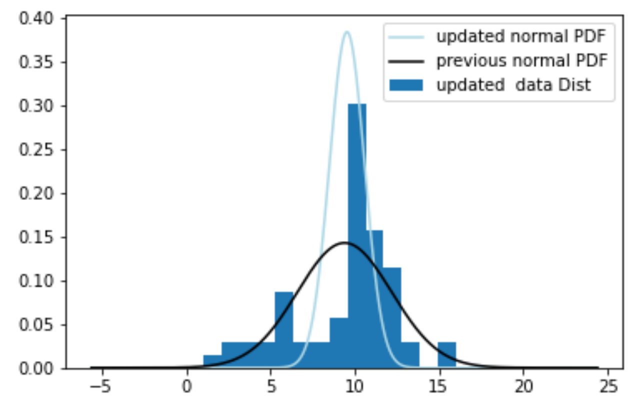


شکل ۶ – تابع چگالی احتمال توزیع $N(9.55, 1.04)$

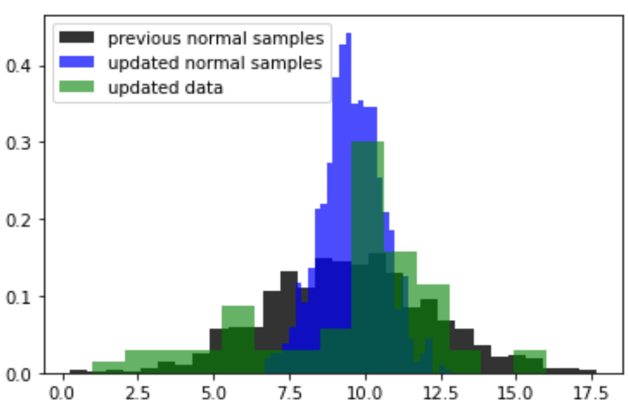


شکل ۷ – هیستوگرام داده‌های تولید شده

۴. در این قسمت به دنبال مقایسه داده‌های تصادفی تولید شده با پارامتر های جدید و اعداد تصادفی تولید شده در قسمت قبل و کل داده‌های موجود از کشتی‌گیر هستیم.



شکل ۸ – مقایسهٔ توزیع‌های نرمال با کل داده‌ها



شکل ۹ – مقایسه توزیع داده‌های تولید شده و کل داده‌ها

با توجه به شکل‌های ۸ و ۹ و پارامترهای توزیع جدید می‌توان دریافت که توزیع تخمین زده شده در این بخش پراکندگی (واریانس) خیلی کمتری نسبت به توزیع بخش قبل دارد. کمتر بودن واریانس موجب افزایش انطباق توزیع تخمین زده شده و توزیع کل داده‌ها در منطقه‌ی محتمل‌تر شده است؛ اما درعین حال در ناحیهٔ دم‌ها که احتمال وقوع کمتر است شاهد عدم انطباق زیادی هستیم.

در این قسمت نیز از مقایسهٔ توزیع داده‌های کشتی‌گیر و نمونه‌های تولید شده می‌توان دریافت که دادّه‌های کشتی‌گیر دارای چولگی هستند در حالی‌که توزیع‌های تخمین زده شده متقارنند.

۵. تنها تفاوت دو مدل عمدتاً در واریانس آن‌هاست. توزیع به‌دست آمده در این بخش واریانس کمتری دارد و به همین خاطر این توزیع، احتمال رخداد داده‌های موجود در دم‌ها را نادیده می‌گیرد (این داده‌ها را کمتر مدل می‌کند). از آن‌جایی که بخش‌های کناری توزیع (به سمت دم‌ها) دارای تعداد قابل توجهی داده هستند، بنابرین پیشرفتی در ارائهٔ توزیع رخ نداده است. همچنین هنوز مشکلاتی همچون متقارن بودن توزیع و عدم انطباق آن در نواحی‌ای که چگالی داده‌ها بالاست، وجود دارد.

مرحلهٔ سوم: در این قسمت یک متغیر تصادفی نشانگر تعریف کردیم که آماده بودن یا نبودن کشتی‌گیر در مسابقات را نشان می‌دهد.

$$Z_i = \begin{cases} 1 & \text{کشتی‌گیر آمادگی داشته باشد} \\ 0 & \text{در غیر این‌صورت} \end{cases}$$

$$Z_i \overset{iid}{\sim} B(1,p)$$
$$Y_i|Z_i,\underline{\theta},\lambda \sim N(\theta_{Z_i},\lambda)$$
$$\theta_{Z_i}|Z_i \sim N(\mu_{Z_i},\sigma^2)$$
$$\lambda = 8.4$$

فرض می‌کنیم قبل از شروع مسابقات جهانی از وضعیت آمادگی این کشتی‌گیر برای هر مسابقه آگاه هستیم. یعنی مقادیر $\underline{z} = (z_1, z_2, \dots, z_n)$ را داریم. حال می‌خواهیم با استفاده از این دادهٔ جدید مدل قبلی را به‌روز رسانی کنیم. یعنی به دنبال پیدا کردن توزیع $(\theta_0,\theta_1)|\ \underline{Z},\underline{Y}$ هستیم.

$$f_{(\theta_0,\theta_1)|\ \underline{Z},\underline{Y}}(u,w) = f_{\theta_0|\ \underline{Z},\underline{Y}}(u) \times f_{\theta_1|\ \underline{Z},\underline{Y}}(w)$$

ابتدا $f_{\theta_0|\ \underline{Z},\underline{Y}}(u)$ را محاسبه می‌کنیم:

$$f_{\theta_0|\ \underline{Z},\underline{Y}}(u) = \frac{f_{\theta_0,\underline{Z},\underline{Y}}\left(u,\underline{z},\underline{y}\right)}{f_{\underline{Z},\underline{Y}}\left(\underline{z},\underline{y}\right)} = \frac{f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right) \times f_{\theta_0|\ \underline{Z}}(u) \times f_{\underline{Z}}\left(\underline{z}\right)}{\int_{-\infty}^{\infty} f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right) \times f_{\theta_0|\ \underline{Z}}(u) \times f_{\underline{Z}}\left(\underline{z}\right) du}$$

$$\rightarrow\ f_{\theta_0|\,\underline{Z},\underline{Y}}(u)=\frac{f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right)\ \times f_{\theta_0|\,\underline{Z}}(u)}{\int_{-\infty}^{\infty} f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right)\ \times f_{\theta_0|\,\underline{Z}}(u)du}$$

ابتدا صورت کسر را محاسبه می‌کنیم:

$$f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right)\ \times f_{\theta_0|\,\underline{Z}}(u)=\prod_{i=1}^{n} f_{Y_i|\,\theta_0=u,\,Z_i}(y_i)\times f_{\theta_0|\,Z_i}(u)$$

$$=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{1}{2}(y_i-u)^2}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(u-\mu_0)^2}$$

برای مخرج کسر داریم:

$$\int_{-\infty}^{\infty} f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right)\ \times f_{\theta_0|\,\underline{Z}}(u)du=\ \int_{-\infty}^{\infty}\prod_{i=1}^{n} f_{\,y_i|z_i,\theta_0=u}(y_i)\times f_{\theta|Z_i=0}(u)du$$

$$=\prod_{i=1}^{n}\int\frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{1}{2}(y_i-u)^2}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(u-\mu_0)^2}\ du$$

$$=\int\prod_{i=1}^{n}\frac{1}{2\pi\sqrt{\lambda\sigma^2}}e^{-\frac{1}{2}(y_i-u)^2}\times e^{-\frac{1}{2}(u-\mu_0)^2}du$$

$$=\frac{1}{2\pi\sqrt{\lambda\sigma^2}}^{n}\int\exp\left\{-nu^2+\left(\sum_{i=1}^{n}y_i+n\mu_0\right)u-\frac{\sum_{i=1}^{n}{y_i}^2}{2}-\frac{n{\mu_0}^2}{2}\right\}du$$

$$A=\ \textstyle\sum_{i=1}^{n}y_i+n\mu_0>0$$

$$B\ =\frac{\sum_{i=1}^{n}{y_i}^2}{2}+\frac{n{\mu_0}^2}{2}>0$$

$$\int e^{-nu^2+Au-B}\,du\ =\ \sqrt{\frac{\pi}{n}}e^{-\frac{4Bn-A^2}{4n}}$$

$$f_{\theta_0|\,\underline{Z},\underline{Y}}(u)=\frac{f_{\underline{Y}|\theta_0,\underline{Z}}\left(\underline{y}\right)\times f_{\theta_0|\,\underline{Z}}(u)}{f_{\underline{Y}|\underline{Z}}\left(\underline{y}\right)}$$

$$=\frac{\exp\left\{-nu^2+\left(\sum_{i=1}^{n}y_i+n\mu_0\right)u-\frac{\sum_{i=1}^{n}{y_i}^2}{2}-\frac{n{\mu_0}^2}{2}\right\}}{\sqrt{\frac{\pi}{n}}e^{-\frac{4Bn-A^2}{4n}}}$$

برای به‌دست آوردن $f_{\theta_1|\,\underline{Z},\underline{Y}}(u)$ نیز مشابه آن‌چه انجام دادیم عمل می‌کنیم:

$$C=\sum_{i=1}^{n}y_i+n\mu_1>0$$
$$D=\frac{\sum_{i=1}^{n}{y_i}^2}{2}+\frac{n{\mu_1}^2}{2}>0$$

$$f_{\theta_1|\,\underline{Z},\underline{Y}}(w)=\frac{f_{\underline{Y}|\theta_1,\underline{Z}}\left(\underline{y}\right)\times f_{\theta_1|\,\underline{Z}}(w)}{f_{\underline{Y}|\underline{Z}}\left(\underline{y}\right)}$$

$$=\frac{exp\left\{-nw^2+\left(\sum_{i=1}^{n}y_i+n\mu_1\right)w-\frac{\sum_{i=1}^{n}{y_i}^2}{2}-\frac{n{\mu_1}^2}{2}\right\}}{\sqrt{\frac{\pi}{n}}e^{-\frac{4Dn-C^2}{4n}}}$$

در نتیجه:

$$\rightarrow f_{\theta_0\theta_1|\,\underline{Z},\underline{Y}}(u,w)$$

$$=\frac{exp\left\{-nw^2+\left(\sum_{i=1}^{n}y_i+n\mu_1\right)w-\frac{\sum_{i=1}^{n}{y_i}^2}{2}-\frac{n{\mu_1}^2}{2}\right\}}{\sqrt{\frac{\pi}{n}}e^{-\frac{4Dn-C^2}{4n}}}$$

$$\times\frac{exp\left\{-nu^2+\left(\sum_{i=1}^{n}y_i+n\mu_0\right)u-\frac{\sum_{i=1}^{n}{y_i}^2}{2}-\frac{n{\mu_0}^2}{2}\right\}}{\sqrt{\frac{\pi}{n}}e^{-\frac{4Bn-A^2}{4n}}}$$