Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Marzieh Alidadi_810101236 Writing Assignment II Deadline 1402/01/31

۱ درست یا غلط؟

۱-۱ زیربخش ۱

این جمله صحیح است.
با استفاده از فرمول‌های زیر می‌توان چارک‌های اول و سوم را محاسبه کرد:

$$Q1 = \mu - (0.675)\sigma \tag{۱}$$

$$Q3 = \mu + (0.675)\sigma \tag{۲}$$

همان‌طور که از این فرمول‌ها مشخص است، این دو چارک، کم‌تر از یک انحراف معیار، از میانگین فاصله دارند.
اگر بخواهیم دقیق‌تر بیان کنیم، این جمله برای اکثر توزیع‌های نرمال برقرار است. اما وجود دارند توزیع‌های نرمالی که این جمله برای آن‌ها صادق نیست و Q۱ و Q۳ آن‌ها بیش از یک انحراف معیار، از میانگین فاصله دارد. در برخی توزیع‌های نرمال که پراکندگی داده‌ها بسیار زیاد است، و یا outlier هایی با فاصله‌های خیلی زیاد در داده‌ها وجود دارد، Q۱ و Q۳ در فاصله‌ی دورتری از میانگین قرار می‌گیرند. و احتمال این وجود دارد که در فاصله‌ی بیش از یک انحراف معیار از میانگین قرار گیرند و جمله‌ی ذکر شده، برای آن‌ها غلط باشد.

۱-۲ زیربخش ۲

این جمله غلط است.
برای یک متغیر تصادفی، شروطی لازم است تا بتوان آن را binomial نامید:
اولاً، n آزمایش مستقل روی نمونه انجام شود.
دوماً، هر آزمایش، یکی از دو نتیجه‌ی موفقیت یا شکست را داشته باشد.
سوماً، احتمال موفقیت در تمام آزمایش‌ها ثابت بماند.
و درنهایت، آزمایش‌ها مستقل از هم باشند.
این‌جا اگر فرض کنیم که فقط دو رنگ موی ممکن وجود دارد، و احتمال مشاهده‌ی هر کدام از آن‌ها بین افراد ثابت است و از دیگران مستقل است، در این‌صورت می‌توان گفت که تعداد افراد با یک رنگ موی خاص، یک متغیر تصادفی binomial است.
که البته این شرایط، در حالت کلی برقرار نیستند و در نتیجه نمی‌توان رنگ موی افراد مورد آزمایش را با یک متغیر تصادفی binomial مدل کرد.

۱-۳ زیربخش ۳

این جمله غلط است.
میانگین و واریانس در توزیع پواسون، همواره با هم برابرند و برابر با پارامتر λ هستند.

۱-۴ زیربخش ۴

این جمله صحیح است.
برای اثبات متقارن بودن یک توزیع، باید عبارت زیر را اثبات کرد:

$$P(X=k) = P(X=n-k) \qquad \forall k \in \{0, ..., n\} \tag{۳}$$

طرفین عبارت بالا، به شکل زیر گسترده می‌شوند:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{n-k} p^k (1-p)^{n-k} \tag{۴}$$

$$P(X = n-k) = \binom{n}{n-k} p^{n-k}(1-p)^k \quad (۵)$$

پس باید اثبات شود که:

$$p^k(1-p)^{n-k} = p^{n-k}(1-p)^k \quad (۶)$$

با توجه به این‌که احتمال موفقیت برابر ۰.۵ است، داریم:

$$\begin{aligned} p = 0.5 \Rightarrow 1-p = p \Rightarrow (1-p)^{n-2k} = p^{n-2k} \\ \Rightarrow (1-p)^{n-k}.(1-p)^{-k} = p^{n-k}.p^{-k} \\ \Rightarrow (1-p)^{n-k}.p^k = p^{n-k}.(1-p)^k \end{aligned} \quad (۷)$$

در نتیجه، اثبات شد که اگر احتمال دوجمله‌ای موفقیت حدوداً برابر ۰.۵ باشد، توزیع تقریباً متقارن است.

۱-۵ زیربخش ۵

این جمله صحیح است.
همان‌طور که در بخش‌های قبل گفته شد، برای یک متغیر تصادفی، شروطی لازم است تا بتوان آن را binomial نامید:
اولاً، n آزمایش مستقل روی نمونه انجام شود.
دوماً، هر آزمایش، یکی از دو نتیجه‌ی موفقیت یا شکست را داشته باشد.
سوماً، احتمال موفقیت در تمام آزمایش‌ها ثابت بماند.
و درنهایت، آزمایش‌ها مستقل از هم باشند.
اینجا، احتمال تعداد موفقیت (تعداد جواب‌های درست)، در یک تعداد خاص آزمایش (تعداد سوالات امتحان)، با احتمال موفقیت مشخص در هر آزمایش (احتمال true یا false بودن برابر ۰.۵)، مورد نظر است. درنتیجه، احتمال بیان شده را می‌توان با یک توزیع binomial مدل کرد.

۲ استنتاج بیزی

برای محاسبه‌ی احتمال خواسته شده، باید احتمال زیر را محاسبه کرد:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{۸}$$

احتمالات استفاده شده در عبارت بالا، عبارتند از:
- p(A|B) : احتمال بهتر بودن مدل جدید، از بهترین مدل موجود، به شرط این‌که مدل جدید، از بین ۲۰ جمله، ۱۵ مورد را به درستی دسته‌بندی کرده‌باشد.

- p(B|A) : احتمال این‌که مدل جدید، از بین ۲۰ جمله، ۱۵ مورد را به درستی دسته‌بندی کند، به شرط بهتر بودن مدل جدید، از بهترین مدل موجود:
این احتمال، به کمک احتمال دو جمله‌ای به شکل زیر قابل محاسبه است:

$$\begin{aligned} P(B|A) &= \binom{20}{15} \cdot (\frac{2}{3} * 0.6)^{15} \cdot (1 - (\frac{2}{3} * 0.6))^5 = 15504 * (0.4)^{15} * (0.6)^5 \\ &= 15504 * 0.000001 * 0.077 = 0.001 \end{aligned} \tag{۹}$$

بخش اول عبارت، تعداد روش‌های ممکن برای انتخاب ۱۵ جمله‌ی درست از بین کل ۲۰ جمله را بیان می‌کند. با توجه به این‌که فرض شده که مدل جدید بهتر از مدل موجود عمل می‌کند، پس حالتی در نظر گرفته شده که در ۲/۳ مواقع، ٪۶۰ از جمله‌ها به درستی دسته‌بندی شوند. بنابراین، احتمال درست دسته‌بندی کردن ۱۵ جمله، در حالتی که مدل جدید بهتر باشد، در بخش دوم عبارت محاسبه شده، و احتمال درست دسته‌بندی نکردن بقیه‌ی ۵ جمله، در بخش سوم محاسبه شده است.

- p(A) : احتمال بهتر بودن مدل جدید، از بهترین مدل موجود:
با توجه به این‌که بیان شده که مدل موجود، ۵۰ درصد جملات را به درستی دسته‌بندی می‌کند، و مدل جدید با احتمال ۱/۳، ۵۰ درصد جملات، و با احتمال ۲/۳، ۶۰ درصد جملات را به درستی دسته‌بندی می‌کند، احتمال بهتر بودن مدل جدید از بهترین مدل

موجود، برابر است با:

$$P(A) = \frac{2}{3} \tag{۱۰}$$

- p(B) : احتمال این‌که مدل جدید، از بین ۲۰ جمله، ۱۵ مورد را به درستی دسته‌بندی کند:
این احتمال به فرم زیر بسط داده می‌شود:

$$P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c) \tag{۱۱}$$

احتمال محاسبه شده در عبارت زیر، احتمال این‌که مدل جدید، از بین ۲۰ جمله، ۱۵ مورد را به درستی دسته‌بندی کند، به شرط بهتر نبودن مدل جدید، از بهترین مدل موجود را بیان می‌کند. در صورتی که مدل جدید بهتر از مدل موجود عمل نکند، یعنی همواره ۵۰ درصد از جمله‌ها را به درستی دسته بندی می‌کند. این احتمال، به فرم زیر قابل محاسبه است:

$$P(B|A^c) = 0.5^{15} = 0.00003 \tag{۱۲}$$

احتمال محاسبه شده در عبارت زیر، احتمال بهتر نبودن مدل جدید، از بهترین مدل موجود را بیان می‌کند. مجموع احتمال بهتر بودن و نبودن آن، برابر ۱ است. احتمال بهتر بودن آن، بالاتر محاسبه شده است. در نتیجه احتمال بهتر نبودن مدل جدید، به فرم زیر قابل محاسبه است:

$$P(A^c) = 1 - P(A) = 1 - \frac{2}{3} = \frac{1}{3} \tag{۱۳}$$

پس احتمال P(B) به فرم زیر قابل محاسبه است:

$$\begin{aligned} P(B) &= P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c) = 0.001 * \frac{2}{3} + 0.00003 * \frac{1}{3} \\ &= 0.0006 + 0.00001 = 0.00061 \end{aligned} \tag{۱۴}$$

در نتیجه، احتمال خواسته شده در سوال، به صورت زیر محاسبه می‌شود:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} = \frac{0.001 * \frac{2}{3}}{0.00061} = \frac{0.0006}{0.00061} = 0.983 = 98.3\% \qquad (۱۵)$$

احتمال بهتر بودن مدل جدید، از بهترین مدل موجود، به شرط اینکه مدل جدید، از بین ۲۰ جمله، ۱۵ مورد را به درستی دسته‌بندی کرده‌باشد، برابر ۹۸.۳ درصد است.

۳ توزیع دوجمله‌ای

۳-۱ زیربخش ۱

با توجه به توضیحاتی که درباره‌ی توزیع binomial در بخش‌های قبل داده شده‌است، تعداد نظرات مثبت، دارای توزیع binomial (دوجمله‌ای) با پارامترهای n=۷۵ و p=۰.۸ است. امید ریاضی این توزیع، به شکل زیر محاسبه می‌شود:

$$E(X) = n \cdot p = 75 * 0.8 = 60 \qquad (۱۶)$$

پس، عدد مورد انتظار برای تعداد نظرات مثبت، برابر ۶۰ است.

۳-۲ زیربخش ۲

اینجا، تعداد موفقیت (تعداد دیدگاه‌های منفی)، در یک تعداد خاص آزمایش مستقل (تعداد ۷۵ دیدگاه مستقل)، با احتمال موفقیت مشخص در هر آزمایش (احتمال دیدگاه منفی برابر ۰.۲)، مورد نظر است. درنتیجه، تعداد دیدگاه‌های منفی را می‌توان با یک توزیع binomial (دوجمله‌ای) مدل کرد.

با توجه به این‌که تعداد ۷۵ دیدگاه، مورد بررسی قرار گرفته است و هر کدام، یکی از دو نوع منفی یا مثبت می‌تواند باشد، پس n=۷۵ در توزیع دوجمله‌ای را داریم. و با توجه به این‌که احتمال منفی بودن یک دیدگاه برابر ۰.۲ و یکسان برای تمام دیدگاه‌ها بیان شده است، پس p=۰.۲ در توزیع دوجمله‌ای را داریم.

پس پارامترهای توزیع دوجمله‌ای بیان شده، برابر (۷۵، ۰.۲) است.

۴ توزیع نمایی

۴-۱ زیربخش ۱

برای بدست آوردن این احتمال، باید از تابع توزیع تجمعی (CDF) استفاده شود. بدین شکل، که سطح زیر نمودار توزیع، از ۰ تا ۲ با هم جمع زده می‌شود، تا احتمال رسیدن بسته در کم‌تر از ۲ روز، محاسبه شود:

$$P(X < x) = 1 - e^{-mx} \tag{۱۷}$$

در این فرمول، مقدار x، برابر ۲ است. و پارامتر m برابر عکس میانگین است:

$$m = \frac{1}{\mu} = \frac{1}{3} \tag{۱۸}$$

پس احتمال خواسته شده، به شرح زیر محاسبه می‌شود:

$$P(X < 2) = 1 - e^{-(\frac{1}{3})2} = 1 - e^{-\frac{2}{3}} \simeq 1 - 0.513 = 0.487 \simeq 0.49 = 49\% \tag{۱۹}$$

در نتیجه، احتمال این‌که بسته در کم‌تر از ۲ روز برسد، حدوداً برابر ۴۹ درصد است.

۴-۲ زیربخش ۲

در این بخش نیز از فرمول استفاده شده در بخش قبل استفاده می‌کنیم؛ با این تفاوت که، سطح زیر نمودار به جای این‌که قسمت سمت چپ نمودار توزیع احتمال تا ۷ مدنظر باشد، این‌بار قسمت سمت راست ۷ در توزیع احتمال مدنظر است:

$$P(X > x) = 1 - P(X < x) = 1 - (1 - e^{-mx}) = 1 - 1 + e^{-mx} = e^{-mx} \tag{۲۰}$$

در این فرمول، مقدار x، برابر ۷ است. و پارامتر m برابر عکس میانگین است:

$$m = \frac{1}{\mu} = \frac{1}{5} \tag{۲۱}$$

پس احتمال خواسته شده، به شرح زیر محاسبه می‌شود:

$$P(X > 7) = e^{-(\frac{1}{5})7} = e^{-\frac{7}{5}} = 0.246 \simeq 0.25 = 25\% \tag{۲۲}$$

در نتیجه، احتمال این‌که بسته در بیشتر از ۷ روز برسد، حدوداً برابر ۲۵ درصد است.

۵ تقریب پواسون

۵-۱ زیربخش ۱

از توزیع پواسون برای توصیف رویدادهای نادر، که احتمال وقوع بسیار پایینی دارند، استفاده می‌شود. برای مثال، احتمال وجود خطا در املای کلمات، در این سوال، از توزیع پواسون پیروی می‌کند. بنابراین، تابع توزیع احتمال برای این رویداد، به شکل زیر در نظر گرفته می‌شود:

$$P(X = k) = \frac{(\lambda^k \cdot e^{-\lambda})}{k!} \tag{۲۳}$$

در این فرمول، با قرار دادن k=۰، توزیع را برای کلمات بدون غلط املایی (تعداد ۰ غلط املایی) در نظر می‌گیریم. پارامتر λ نیز در توزیع پواسون، برابر با میانگین و انحراف معیار است. با توجه به این‌که ۰.۵ درصد کلمات دارای غلط املایی هستند، احتمال وجود غلط املایی، برابر ۰.۰۰۵ است. بنابراین، با توجه به این‌که این احتمال برای هر صفحه نیر صادق است، و همچنین، ۱۰۰ کلمه در هر صفحه وجود دارد، میانگین تعداد غلط املایی در هر صفحه، به شرح زیر محاسبه می‌شود و از آن برای محاسبه‌ی λ استفاده می‌شود:

$$\begin{aligned}\lambda &= mean\ of\ the\ number\ of\ words\ having\ spell\ errors\ per\ image\\ &= probability\ of\ words\ having\ spell\ errors * number\ of\ words\ per\ image\\ &= 0.005 * 100 = 0.5\end{aligned} \tag{۲۴}$$

پس احتمال نظیر وجود کلمات بدون غلط املایی، به شرح زیر محاسبه می‌شود:

$$P(X=0) = \frac{(0.5^0 * e^{-0.5})}{0!} == \frac{1 * 0.606}{1} = 0.606 \simeq 0.61 = 61\% \tag{۲۵}$$

در نتیجه، حدوداً ۶۱ درصد از کلمات، بدون غلط املایی خواهند بود.

۵-۲ زیربخش ۲

برای محاسبه‌ی احتمال وجود کلمات با ۲ یا بیشتر غلط املایی، از تابع توزیع تجمعی (CDF) پواسون با پارامتر λ که در بخش قبل محاسبه شد، استفاده می‌کنیم:

$$\begin{aligned} P(X >= 2) &= 1 - P(X < 2) = 1 - P(X <= 1) = 1 - \sum_{k=0}^{1} \frac{(\lambda^k \cdot e^{-\lambda})}{k!} \\ &= 1 - (\frac{(0.5^0 * e^{-0.5})}{0!} + \frac{(0.5^1 * e^{-0.5})}{1!}) = 1 - \frac{(1 * e^{-0.5})}{1} - \frac{(0.5 * e^{-0.5})}{1} \\ &= 1 - (e^{-0.5}) - (0.5 * e^{-0.5}) = 1 - 1.5 * e^{-0.5} = 1 - 1.5 * 0.606 \\ &= 1 - 0.909 = 0.091 \simeq 0.09 = 9\% \end{aligned} \tag{۲۶}$$

در نتیجه، حدوداً ۹ درصد از کلمات، ۲ یا بیشتر غلط املایی خواهند داشت.

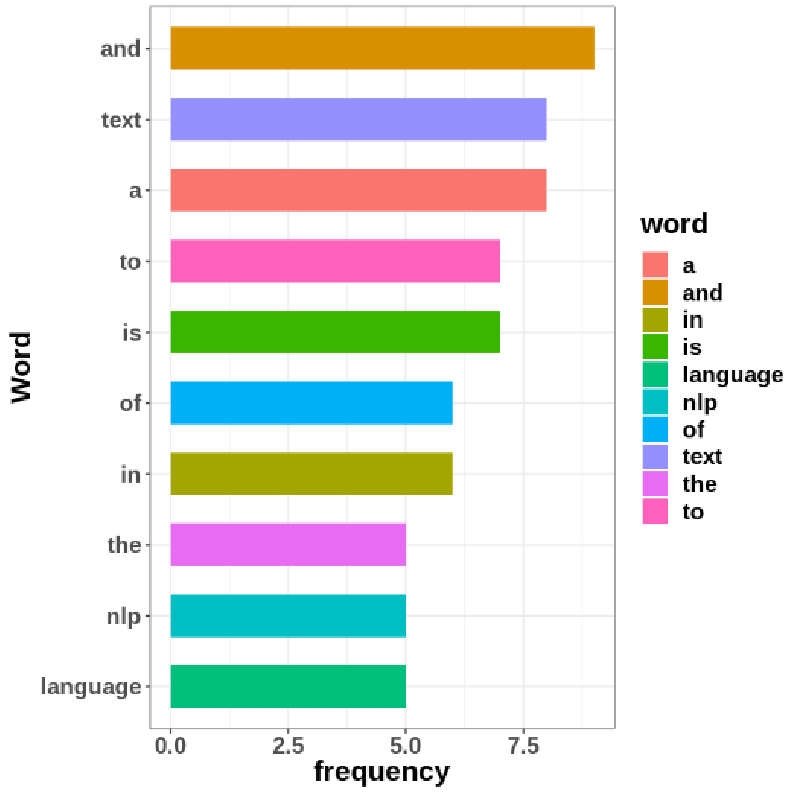
۶ برنامه‌نویسی R

۶-۱ زیربخش ۱

۱۰ کلمه با بیشترین تعداد تکرار در متن داده شده، و تعداد تکرار هر یک، به شرح زیر است:

```
        word frequency
1        and         9
2          a         8
3       text         8
4         is         7
5         to         7
6         in         6
7         of         6
8   language         5
9        nlp         5
10       the         5
```

نمودار bar chart مربوط به این ۱۰ کلمه، در شکل زیر رسم شده‌است:



۶-۲ زیربخش ۲

ابتدا جملات متن داده شده، از هم تفکیک شدند. در کل، ۱۴ جمله در متن وجود داشت: (جملات طولانی، در عکس زیر، به طور کامل نمایش داده نشده‌اند.)

```
 [1] "The power of words cannot be underestimated"
 [2] " Language is a critical component of human communication,\nand natural language processing (NLP) is a field that se
 [3] " NLP involves developing algorithms and computational models that can\nanalyze, interpret, and generate human langu
 [4] " One of the most common tasks in NLP is text classification"
 [5] "\nIn this task, an algorithm is trained to automatically assign predefined categories or labels to a given text"
 [6] "\nFor example, a text classification algorithm may be trained to identify whether an email is spam or not,\nbased c
 [7] " Another important task in NLP is named entity recognition (NER)"
 [8] " NER\ninvolves identifying and classifying named entities in a text, such as people, organizations, and locations"
 [9] "\nThis task is useful in applications such as information extraction and text-to-speech synthesis"
[10] " Other tasks in\nNLP include sentiment analysis, machine translation, and text summarization"
[11] " Sentiment analysis involves\ndetermining the sentiment or emotion expressed in a given text, such as positive or r
[12] " Machine\ntranslation involves automatically translating text from one language to another, while text summarizatic
[13] " In order to perform these tasks, NLP algorithms\ntypically rely on statistical models and machine learning technic
[14] " These techniques involve training the\nalgorithm on large amounts of data, so that it can learn to recognize patte
```

سپس، برای هر یک از جملات، تعداد حروف کلمات حاضر در آن‌ها، به ترتیب، مشخص شد. و در یک dataframe به شکل زیر ذخیره شد:

	sentence	word_length
	<int>	<int>
1	1	3
2	1	5
3	1	2
4	1	5
5	1	6
6	1	2
7	1	14
8	2	8
9	2	2
10	2	1

در نهایت، نمودار جعبه‌ای گروهی به تفکیک جملات، نشان‌دهنده‌ی توزیع تعداد حروف کلمات آن‌ها، به فرم زیر رسم شد: