HW 1

Probability Home Assignment.

Work has been done by: Ryabykin Aleksey

Problem 1. **Prove that**

$$\sum_{j=0}^{k} \binom{n}{j} \binom{m}{k-j} = \binom{m+n}{k}.$$

By that, we are proving that a sum of two binomial random variables $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ with the same success probability p, is also a binomial random variable $X + Y \sim \mathrm{Bin}(n + m, p)$.

Proof: I will prove this statement in an algebraic way. Let me use the binomial theorem:

> **Theorem: (A simple variant of a binomial theorem)**
>
> $$(1+x)^n = \binom{n}{0}x^0 + \binom{n}{1}x^1 + \ldots + \binom{n}{n-1}x^{n-1} + \binom{n}{n}x^n,$$
>
> or equivalently,
>
> $$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

We can rewrite the right and left parts of the identity in terms of binomial theorem:

$$\underbrace{(1+x)^n}_{\sum_{j=0}^{n}\binom{n}{j}x^j} \cdot \underbrace{(1+x)^m}_{\sum_{p=0}^{m}\binom{m}{p}x^p} = \sum_{k=0}^{m+n} \binom{m+n}{k} x^k$$

$$\sum_{j=0}^{n}\sum_{p=0}^{m} \binom{n}{j}\binom{m}{p} x^{j+p}.$$ Let $k = j + p$, then $p = k - j$.

Making this changes in variables we can obtain:

$$\sum_{j=0}^{n}\sum_{k=j}^{j+m} \binom{n}{j}\binom{m}{k-j} x^k$$

Since $j \leq n$ we can split the sums:

$$\sum_{j=0}^{n}\sum_{k=j}^{n+m} \binom{n}{j}\binom{m}{k-j} x^k - \cancel{\sum_{j=0}^{n}\sum_{k=j+m+1}^{n+m} \binom{n}{j}\binom{m}{k-j} x^k}$$

We can reduce the last one term because of the next thoughts. We know, that $k > j + m$ (easily get it from the limits), so the binomial coefficient $\binom{n}{r-k}$ will be equal to zero. Similarly,

$$\sum_{j=0}^{n}\sum_{k=0}^{n+m}\binom{n}{j}\binom{m}{k-j}x^k - \cancel{\sum_{j=0}^{n}\sum_{k=0}^{k-1}\binom{n}{j}\binom{m}{k-j}x^k}$$

Reducing is legal because of $k < j$, then binomial coefficient $\binom{m}{k-j}$ is equal to 0. So, we can change sum operators:

$$\sum_{k=0}^{n+m}\sum_{j=0}^{n}\binom{n}{j}\binom{m}{k-j}x^k$$

In case of $k \geq n$ we are getting:

$$\sum_{k=0}^{n+m}\sum_{j=0}^{k}\binom{n}{j}\binom{m}{k-j}x^k - \cancel{\sum_{k=0}^{n+m}\sum_{j=n+1}^{k}\binom{n}{j}\binom{m}{k-j}x^k}$$

Reduced because of $j > n$. Otherwise, if $k < n$:

$$\sum_{k=0}^{n+m}\sum_{j=0}^{k}\binom{n}{j}\binom{m}{k-j}x^k + \cancel{\sum_{k=0}^{n+m}\sum_{j=k+1}^{n}\binom{n}{j}\binom{m}{k-j}x^k}$$

Legally simplifying since $j > k$. Either the 1st case or the 2nd leads us to the:

$$\sum_{k=0}^{n+m}\sum_{j=0}^{k}\binom{n}{j}\binom{m}{k-j}x^k$$

on the left And

$$\sum_{k=0}^{m+n}\binom{m+n}{k}x^k$$

Comparing the coefficient of x^k, we can finally obtain:

$$\sum_{j=0}^{k}\binom{n}{j}\binom{m}{k-j} = \binom{m+n}{k}.$$

□

Problem 2.

A basket contains n balls, out of which w are white and b are black ($w + b = n$). We extract m balls from this basket with replacement and note their colors. Find the probability that out of these m balls exactly k were white.

Solution: Suppose A is an event defined as from m balls have been chosen exactly k white ones. We can obtain the needed probability $P(A)$ by the following formula:

$$P(A) = C_m^k \frac{w^k \cdot b^{m-k}}{n^m},$$

where n^m means all possible variations, $w^k \cdot b^{m-k}$ mean k white balls has been extracted with the rest $m-k$ black balls. And C_m^k is a number of combinations, which can be extracted m balls from n ones. ■

Problem 3.

A fair dice is rolled n times. What is the probability that at least 1 of the 6 values never appears?

Solution: Using inclusion/exclusion formula the probability would be:

$$P = \binom{6}{1}\left(\frac{5}{6}\right)^n - \binom{6}{2}\left(\frac{4}{6}\right)n + \binom{6}{3}\left(\frac{3}{6}\right)^n - \binom{6}{4}\left(\frac{2}{6}\right)^n + \binom{6}{5}\left(\frac{1}{6}\right)^n$$

■

Problem 4.

There $(m+1)$ baskets and each basket has exactly m balls in it. Additionally for each $n = 0, 1, \dots, m$, we know that n-th basket contains exactly n white and $(m-n)$ black balls. We pick a basket at random and pick k balls from it with replacement. Find the probability that $(k+1)$-th ball will be white if all k balls were white.

Solution: The core idea is that probability of extracting a white ball at any timestamp is the same because of with replacement case. That is why, for example, for basket i:

$$P_{k+1} = \frac{n}{m},$$

where n is a number of white balls in the basket i, m – all balls in the basket. Now we can use the law of total probability, it will be equal to:

$$P = \sum_{n=1}^{m} P(B_i)P(\text{white at } k+1 \mid k \text{ balls were white}) = \sum_{n=1}^{m} \frac{1}{m+1}\frac{n}{m}.$$

Iterating through baskets with just counting the choosing this basket and a probability to choose a white ball on $k+1$st step with chosen k whites. ■

Problem 5.

Let $X_1,\ X_2 \overset{i.i.d.}{\sim} U(0,1)$. Find the pdf of $Z = X_1 \cdot X_2$.

Solution: We can find the distribution function:

$$P(Z \leq z) = \int\limits_0^1 P(xY \leq z) f_X(x) dx = \int\limits_0^1 P\left(Y \leq \frac{z}{x}\right) f_X(x) dx.$$

Splitting the integral:

$$P(Z \leq z) = \int\limits_0^z f_X(x) dx + \int\limits_z^1 P\left(Y \leq \frac{z}{x}\right) f_X(x) dx =$$

$$= \int\limits_0^z dx + \int\limits_z^1 \frac{z}{x} dx = z - 0 + 0 - z\log z = z - z\log z = F_Z(z)$$

Hence the pdf of Z,

$$f_Z(z) = \frac{d}{dz} F_Z(z) = 1 - \left(\log z + \frac{z}{z}\right) = -\log z.$$

■

Problem 6.

If someone gets a positive result on a COVID test that only gives a false positive with probability 0.001, what is the chance that he or she actually got COVID, if:

1. The probability that a person has COVID is 0.01;
2. The probablity that a person has COVID is 0.0001.

Solution:

1. Let me define some notations. Let define $P(\text{COVID})$ as the probability that a person has COVID, $P(\text{HEALTHY})$ as the probability that a person does not have a COVID, $P(\text{POS})$, $P(\text{NEG})$ as probabilites of positive and negative tests results respectively. Then to obtain the probability of actually be COVID illed when a patient got positive result we can imply Bayes' theorem:

   $$P(\text{COVID} \mid \text{POS}) = \frac{P(\text{POS} \mid \text{COVID}) P(\text{COVID})}{P(\text{POS})}.$$

   We already know $P(\text{POS} \mid \text{COVID}) = 1$, $P(\text{COVID}) = 0.01$. The only thing we need is the probability of positive result, that can be found by the law of total probability:

   $$P(\text{POS}) = P(\text{POSITIVE} \mid \text{COVID}) P(\text{COVID}) + P(\text{POSITIVE} \mid \text{HEALTHY}) P(\text{HEALTHY}) =$$
   $$= 1 \cdot 0.01 + 0.001 \cdot 0.99 = 0.01099$$

   So, the final probability for the first case:

   $$P(\text{COVID} \mid \text{POS}) = 0.91.$$

2. Similarly, lets apply the same formulas:

   $$P(\text{POS}) = P(\text{POSITIVE} \mid \text{COVID}) P(\text{COVID}) + P(\text{POSITIVE} \mid \text{HEALTHY}) P(\text{HEALTHY}) =$$
   $$= 1 \cdot 0.0001 + 0.001 \cdot 0.99 = 0.00109$$

   And, finally

   $$P(\text{COVID} \mid \text{POS}) = \frac{1 \cdot 0.0001}{0.00109} = 0.0917.$$

■

Problem 7.

Let X be normally distributed $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, so $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Find the PDF of $Y = X^2$.

Solution: We can obtain the distribution function by the followng steps:

$$F_Y(y) = P(Y \le y) = P\left(X^2 \le y\right) = P\left(|X| \le \sqrt{y}\right) = P\left(-\sqrt{y} \le X \le \sqrt{y}\right)$$

That can be easily rewritten in the form:

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

To finally obtain the probability density function we can differentiate the last expression:

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y})$$

Keeping in mind evenness of Gaussian function, we can just simplify:

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}) = \frac{1}{\sqrt{y2\pi\sigma^2}} \exp\left(-\frac{(\sqrt{y}-\mu)^2}{2\sigma^2}\right).$$

■

Problem 8.

Find $\mathbb{E}\left(|X-Y|\right)$ for $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

Solution: We know, that any linear combination of normally distributed random variables also have a normal distribution. We can express expectation and standard deviation of the non-absolute value of given difference just by knowledge of this fact:

$$\begin{gathered} W = X - Y \sim \mathcal{N}(0,2) \\ \mu_W = 0 - 0 = 0 \\ \sigma_W^2 = \sigma_X^2 + (-1)^2\sigma_Y^2 = 1 + 1 = 2. \end{gathered}$$

Now let's find the expectation of an absolute value:

$$\mathbb{E}|W| = \int_{-\infty}^{\infty} |w| \frac{1}{\sqrt{2\pi \cdot 2}} \exp\left(-\frac{(w-0)^2}{2\cdot 2}\right)$$

$$\mathbb{E}|W| = \int_{-\infty}^{\infty} |w| \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{w^2}{4}\right) = -\frac{1}{\sqrt{\pi}} \cdot \int_{-\infty}^{0} \frac{w}{2} \exp\left(-\frac{w^2}{4}\right) + \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \frac{w}{2} \exp\left(-\frac{w^2}{4}\right) =$$

$$= \left| \begin{matrix} \text{Let } u = \frac{w^2}{4} \\ du = \frac{w}{2} \end{matrix} \right| =$$

$$= \frac{1}{\sqrt{\pi}} e^{-u} \Big|_{-\infty}^{0} - \frac{1}{\sqrt{\pi}} e^{-u} \Big|_{0}^{\infty} = \frac{1}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}} = \frac{2}{\sqrt{\pi}}.$$

■

Problem 9.

We found that the sum of $X_1, X_2 \overset{i.i.d.}{\sim} U(0,b)$, $X_1 + X_2$ has a "triangular" PDF.

1. Find the PDF of $Y = X_1 + X_2 + X_3$ for $X_1, X_2, X_3 \overset{i.i.d.}{\sim} U(0,b)$, where, for now, $b = 1$.
2. Find $\mathbb{E}Y$ and $\operatorname{var} Y$. What happens to them if b is set to something $\neq 1$ (but still $b > 0$ for simplicity).

Solution: 1. As it has been said before, we've found that the sum of two uniform distributed random variables has a "triangular" PDF of a kind:

$$f_T(t) = \begin{cases} t, & 0 < t < 1, \\ 2 - t, & 1 < t < 2, \end{cases}$$

where $T = X_1 + X_2$. So that is why our goal is to find a new sum S equal to $S = T + X_3$, where T is distributed with pdf $f_T(t)$ and X_3 has a uniform distribution. The core idea is to implement convolution again to obtain the answer.

$$f_S(s) = \int\limits_{-\infty}^{+\infty} f_T(s-t) f_S(t) dt.$$

So,

$$f_S(s) = \begin{cases} \int\limits_0^s t dt = \frac{s^2}{2}, & 0 < s < 1, \\ \int\limits_{s-1}^s (2-t) dt = -s^2 + 3s - \frac{3}{2}, & 1 < s < 2, \\ \int\limits_{s-1}^2 (2-t) dt = \frac{(s-3)^2}{2}, & 2 < s < 3. \end{cases}$$

2. Keeping in mind the linearity of expectation:

$$\mathbb{E}Y = \mathbb{E}X_1 + \mathbb{E}X_2 + \mathbb{E}X_3 = 3 \cdot \frac{1}{2}.$$

Suppose b is an arbitrary variable:

$$\mathbb{E}Y = \frac{3b}{2}.$$

Because of independence of random variables X_n $n \in \{1, 2, 3\}$, then:

$$\operatorname{var} Y = \sum_{i=1}^{3} \operatorname{var} X_i = \frac{3}{12}$$

In case when b has an arbitrary value:

$$\operatorname{var} Y = \sum_{i=1}^{3} \operatorname{var} X_i = \frac{3b^2}{12}$$

■

Problem 10.

Work out that, for $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, the magnitude $R = \sqrt{X^2 + Y^2}$ of a random vector (X, Y) is Rayleigh-distributed, $R \sim$ Rayleigh(σ):

$$f_R(r|\sigma) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

<u>Solution:</u> We can standardize a normally distributed random variables. So, $\frac{X}{\sigma}, \frac{Y}{\sigma} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then the magnitude can be written in the following manner:

$$R = \sigma \sqrt{\frac{X^2}{\sigma^2} + \frac{Y^2}{\sigma^2}}.$$

Keeping in mind, that chi-squared distribution with k freedom degrees is a distribution of a sum of squares of k independent identity distributed standard random variables, we can obtain, that $R = \sigma\sqrt{\chi^2(2)}$. Let's write the distribution function:

$$P(R \le r) = P\left(\sigma\sqrt{\chi^2(2)} \le r\right) = P\left(\sqrt{\chi^2(2)} \le \frac{r}{\sigma}\right) = P\left(\chi^2(2) \le \frac{r^2}{\sigma^2}\right) = F_R(r)$$

Hense, probability density function can be obtained by differentiation:

$$f_R(r) = f_{\chi^2(2)}\left(\frac{r^2}{\sigma^2}\right) \cdot \frac{2r}{\sigma^2}.$$

We know the probability density function of $\chi^2(2)$:

$$f(x; k) = \frac{x^{\frac{k}{2} - 1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} = e^{-\frac{x}{2}} \cdot \frac{1}{2}.$$

Finally,

$$f_R(r) = \frac{2r}{\sigma^2} \cdot \frac{1}{2} \cdot \exp\left(-\frac{r^2}{2\sigma^2}\right) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) = f_R(r|\sigma).$$

■

Problem 11.

We will say that set $A \subset \mathbb{N}$ has asymptotic density θ if there exits the following limit:

$$\lim_{n\to\infty} \frac{|A \cap \{1, \ldots, n\}|}{n} = \theta.$$

Denote the family of such sets (which have asymptotic density) as $\mathcal{A}$. Is $\mathcal{A}$ a σ-algebra?

<u>Solution:</u> Let me prove it by the definition of σ-algebra.

Definition: (σ-algebra)

Let X be some set. Then a subset σ of the powerset of the set X is called a σ-algebra if it satisfies the following properties:

1. Either X, or $\varnothing$ are contained in σ;
2. If $E \in \sigma$, then its complement is contained in σ: $X \setminus E \in \sigma$;
3. Union and intersection of countable subsets of σ are contained in σ.

Let it prove step by step:

1. $\varnothing \in \mathcal{A}$:

$$\lim_{n\to\infty} \frac{|\varnothing \cap \{1, \dots, n\}|}{n} = 0$$

Also $\mathbb{N} \in \mathcal{A}$:

$$\lim_{n\to\infty} \frac{|\mathbb{N} \cap \{1, \dots, n\}|}{n} = 1.$$

2. Let $A \in \mathcal{A}$, then $\mathbb{N} \setminus A \in \mathcal{A}$:

$$\frac{|\,(\mathbb{N} \setminus A) \cap \{1, \dots, n\}|}{n} \leq 1$$

Following the Weierstrass theorem, this limit exists.

3. Similarly, for union and intersection of countable subsets of $\mathcal{A}$.

■