Midterm 1: Solutions

Introduction to Data Science

October 5, 2017

1 Probability

1.1 Moment Generating Functions

The moment generating function of a random variable X is

$$M_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tx} f(x) dx \tag{1}$$

and think of f as a continuous random variable, although the result holds for discrete random variables as well.

Question 1: Compute $M_X^{(r)}(t) = \frac{d^r}{dt^r} M_X(t)$. Note that $\frac{d^r}{dt^r} \int_{-\infty}^{+\infty} h(x,t)dx = \int_{-\infty}^{+\infty} \frac{\partial^r h(x,t)}{\partial t^r} dx$

Solution 1:

$$M_X^{(r)}(t) = \frac{d^r}{dt^r} \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \int_{-\infty}^{+\infty} x^r e^{tx} f(x) dx \tag{2}$$

Question 2: Compute $M_X^{(r)}(t=0)$ and express it as an expected value of powers of X.

Solution 2:

$$M_X^{(r)}(0) = \int_{-\infty}^{+\infty} x^r f(x) dx = E\left[X^r\right] \tag{3}$$

Let's try to understand how moment generating functions behave under linear transformations. We first need to establish some facts. If X and Y are 2 independent random variables then:

$$E\left[XY\right] = E\left[X\right] E\left[Y\right] \tag{4}$$

and if a if a constant

$$E\left[aX\right] = aE\left[X\right] \tag{5}$$

Question 3: If X and Y are independent random variables with moment generating functions M_X and M_Y and $Z = X + Y$, compute M_Z as a function of M_X and M_Y.

Solution 3:

$$M_Z(t) = E\left[e^{tZ}\right] = E\left[e^{t(X+Y)}\right] = E\left[e^{tX}e^{tY}\right] = E\left[e^{tX}\right]E\left[e^{tY}\right] = M_X(t)M_Y(t) \tag{6}$$

Question 4: If X is a random variable with moment generating function M_X and $Y = a + bX$, compute M_Y as a function of M_X.

Solution 4:

$$M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(a+bX)}\right] = E\left[e^{ta}e^{tbX}\right] = e^{ta}E\left[e^{tbX}\right] = e^{ta}M_X(tb) \tag{7}$$

Question 5: If X is a random normal variable then its moment generating function is

$$M_X(t) = \exp(t\mu + \frac{1}{2}\sigma^2 t^2) \tag{8}$$

Compute the first and second moment of X

Solution 5: We compute the first and second derivative

$$\begin{aligned} M_X^{(1)}(t) &= \frac{d}{dt}\exp(t\mu + \frac{1}{2}\sigma^2 t^2) \\ &= (\mu + \sigma^2 t)\exp(t\mu + \frac{1}{2}\sigma^2 t^2) \end{aligned} \tag{9}$$

and

$$\begin{aligned} M_X^{(2)}(t) &= \frac{d}{dt}(\mu + \sigma^2 t)\exp(t\mu + \frac{1}{2}\sigma^2 t^2) \\ &= \sigma^2\exp(t\mu + \frac{1}{2}\sigma^2 t^2) + (\mu + \sigma^2 t)^2\exp(t\mu + \frac{1}{2}\sigma^2 t^2) \end{aligned} \tag{10}$$

Therefore $M_X^{(1)}(0) = \mu$ and $M_X^{(2)}(0) = \sigma^2 + \mu^2$

Question 6: If $Var[X] = E[X^2] - E[X]^2$, compute $Var[X]$ for a random normal variable.

Solution 6: $Var[X] = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$

1.2 Disease test

A patient goes to see a doctor. The doctor performs a test with 99 percent reliability–that is, 99 percent of people who are sick test positive and 99 percent of the healthy people test negative. The doctor knows that only 1 percent of the people in the country are sick. Now the question is: if the patient tests positive, what are the chances the patient is sick? Let's break down the problem. A lot have to do with translating the problem correctly.

- T is the event "the test is positive"
- $\bar{T}$ is the event "the test is negative"
- D is the event "the patient has the disease"
- $\bar{D}$ is the event "the patient does not have the disease"

Question 1: What are the probability values of $P(T|D)$, $P(\bar{T}|\bar{D})$, $P(T|\bar{D})$, $P(\bar{T}|D)$, $P(D)$ and $P(\bar{D})$?

Solution 1:

- $P(T|D) = 0.99$
- $P(\bar{T}|\bar{D}) = 0.99$
- $P(T|\bar{D}) = 1 - P(\bar{T}|\bar{D}) = 0.01$
- $P(\bar{T}|D) = 1 - P(T|D) = 0.01$
- $P(D) = 0.01$
- $P(\bar{D}) = 1 - P(D) = 0.99$

If A and B are possible events, the Bayes theorem states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{11}$$

Additionally the law of total probability states that if $\{B_n|n : 1, 2, \ldots, k\}$ is a set of events

$$P(A) = \sum_{n=1}^{k} P(A \cap B_n) = \sum_{n=1}^{k} P(A|B_n)P(B_n) \tag{12}$$

Question 2: Using the Bayes theorem and the law of total probability, compute $P(D|T)$, that is the probability that the patient is sick if the test is positive.

Solution 2: We have

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\ P(D|T) &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} \\ P(D|T) &= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} \\ P(D|T) &= \frac{1}{2} \end{aligned} \tag{13}$$

1.3 Mean and variance of probability distribution

The expected value random variable with a discrete probability distribution is defined by

$$\mu = E[X] = \sum_{x \in \mathcal{X}} xP(x) \tag{14}$$

and in the continuous case

$$E[X] = \int_{x \in \mathcal{X}} x dp(x) \tag{15}$$

The variance is defined as

$$Var[X] = E[(X - \mu)^2] \tag{16}$$

The probability mass function of random variable $X \sim B(n, p)$ with binomial distribution is

$$P(X = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{17}$$

n being the total number of experiments and p the probability of each experiment yielding a successful result.

Question 1: Compute the mean (hint: $\sum_{k=0}^{n} P(k; n, p) = 1$ for any n)

Solution 1: We have

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} kP(k) \\ &= \sum_{k=0}^{n} k\binom{n}{x}p^k(1-p)^{n-k} \\ &= np\sum_{k=0}^{n} k\frac{(n-1)!}{(n-k)!k!}p^{k-1}(1-p)^{n-k} \\ &= np\sum_{k=1}^{n} \frac{(n-1)!}{(n-1-k+1)!(k-1)!}p^{k-1}(1-p)^{n-1-k+1} \\ &= np\sum_{l=0}^{n-1} \frac{(n-1)!}{(n-1-l)!l!}p^{l}(1-p)^{n-1-l} \\ &= np\sum_{l=0}^{m} \frac{m!}{(m-l)!l!}p^{l}(1-p)^{m-l} \\ &= np \end{aligned} \tag{18}$$

Question 2: Prove that $E[(X-\mu)^2] = E[X^2] - E[X]^2$

Solution 2: We have

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2] + \mu^2 - 2E[X\mu] \\ &= E[X^2] + E[X]^2 - 2E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned} \tag{19}$$

Question 3: Compute $E[X(X-1)]$

Solution 3:

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)P(k) \\ &= \sum_{k=0}^{n} k(k-1)\binom{n}{x}p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n} \frac{n!}{(n-k)!(k-2)!}p^k(1-p)^{n-k} \\ &= n(n-1)p^2\sum_{k=2}^{n} \frac{(n-2)!}{(n-k)!(k-2)!}p^{k-2}(1-p)^{n-k} \\ &= n(n-1)p^2\sum_{l=0}^{m} \frac{m!}{(m-l)!l!}p^{l}(1-p)^{m-l} \\ &= n(n-1)p^2 \end{aligned} \tag{20}$$

Question 4: Compute the variance

Solution 4:

$$\begin{aligned} Var[X] &= E[X^2] - E[X]^2 \\ &= E[X(X-1)] + E[X] - E[X]^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \end{aligned} \tag{21}$$

1.4 Car on the road

The probability of observing at least one car on a highway during any 20-minutes time interval is 609/625. Assume that the probability of seeing a car at any moment is uniform (constant) for the entire 20 minutes.

Question 1: What is the probability that we do not observe any car during a 20-minute interval?

Solution 1: $1 - 609/625$

Question 2: We call p the probability of observing a car in any 5-minute interval. What is the probability that we do not observe any car during 4 non-overlapping 5-minute independent intervals?

Solution 2: $(1-p)^4$

Question 3: Using question 1 and 2, compute p.

Solution 3: $(1-p)^4 = 1 - 609/625 = 16/625 \Rightarrow p = 3/5$

2 Statistics

We define the Pearson correlation coefficient $\rho_{X,Y}$ between 2 random variables X and Y to be

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{22}$$

where μ_X, μ_Y, σ_X and σ_Y are the respective mean of standard deviation of X and Y.

Question 1: Suppose X is uniformly distributed on $[0, 2\pi]$. Let $Y = \sin(X)$ and $Z = \cos(X)$. The probability density function of the continuous uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-1} & \text{for } a \leq x \leq b \\ 0 & \text{for } a \geq x \text{ or } x \geq b. \end{cases} \tag{23}$$

Compute the correlation between Y and Z

Solution 1: We have

$$E[(Z-\mu_Z)(Y-\mu_Y)] \quad = \quad E[ZY] - E[Z]E[Y] \tag{24}$$

We have

$$E[Z] = \frac{1}{2\pi}\int_0^{2\pi} \cos(x)dx = 0 \tag{25}$$

$$E[Y] = \frac{1}{2\pi}\int_0^{2\pi} \sin(x)dx = 0 \tag{26}$$

and

$$E[ZY] = \frac{1}{2\pi}\int_0^{2\pi} \sin(x)cos(x)dx = 0 \tag{27}$$

Therefore $\rho_{X,Z} = 0$ even then they are deterministically dependent on each other.

3 Statistical inference

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You are basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment will not be repeatable and so has little use.

3.1 One-Tailed Hypothesis Testing Example

A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of $\sigma = 15$. The IQ scores are normally distributed.

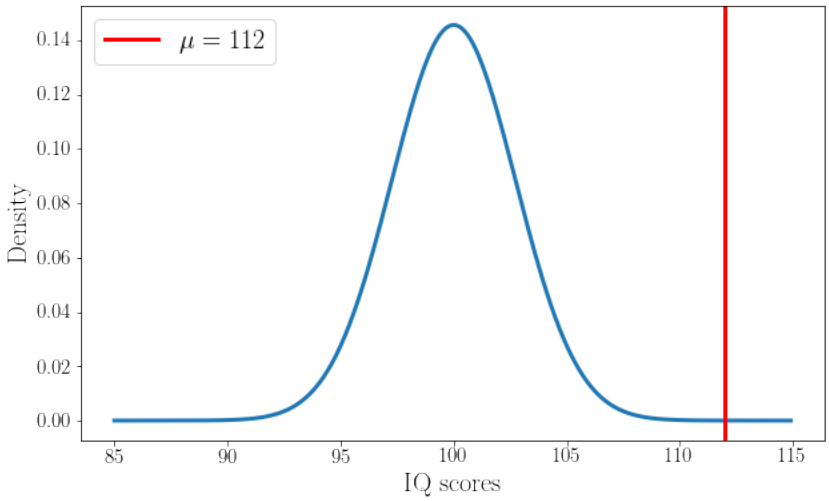
The hypothesis statement in this question is that the principal believes the average intelligence in school is more than 100. It can be written in mathematical terms as:

$$H_1 : \mu > 100 \tag{28}$$

H_1 is called the *Alternate Hypothesis*. This is the claim we need to study. The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test. It would be called two-tailed if $H_1 : \mu \neq 100$. The accepted fact is that the population mean is 100 that we state as

$$H_0 : \mu = 100 \tag{29}$$

H_0 is called the *Null hypothesis*. To visualize, here is the distribution of $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $n = 30$ and X_is are independent random normal variables.



We have $Var[\bar{X}] = \frac{\sigma^2}{n}$. Here $\bar{X} = 112$ but we know that the true mean of n normal distributed variables with mean μ is μ. So we can wonder how far is $\bar{X}$ from its true value μ (if we had sampled an infinite amount of time). The z-score:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{30}$$

measures, in standard deviation units, how far $\bar{X}$ is from μ. Note that somehow in this problem we know the true mean and standard deviation of the population with μ and σ

being parametric constants (we are not estimating them using samples). As a consequence z is a random variable being normal distributed

Question 1: Compute the z-score in this case.

Solution 1: $z = \frac{112-100}{15/\sqrt{30}} \simeq \frac{12}{2.74} \simeq 4.37$

Question 2: Considering the z-score value, do you think it is likely that the average of 112 computed over 30 students indicates that the students in this principal's school are above average intelligence? As a reference $P(z > 0) = 1/2$ and $P(z > 2) = 0.02275$.

Solution 2: It is very likely because the sample average is more than 4 standard deviations above what is expected in average. It is a very unlikely event that we could sample 30 students with IQ $\sim \mathcal{N}(\mu, \sigma)$ and get $\bar{X} = 112$.

3.2 Confidence Interval

The confidence interval is the interval such that the true population statistics is contained with a certain level of confidence between a lower and upper bounds learned from a sample of data. Let's consider $\{X_1, X_2, \ldots, X_n\}$ i.i.d. normal distributed random variables with population mean μ and standard deviation σ. The sample mean is

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{31}$$

and

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{32}$$

is normal distributed with mean $\mu_z = 0$ and standard deviation $\sigma_z = 1$. We now want to find a upper bound c and lower bound $-c$ such that

$$Pr(-c \leq T \leq c) = 1 - \alpha \tag{33}$$

where it is common to take $\alpha = 0.05$ (95% confidence) and we are going to make this assumption from now on. We have

$$\begin{aligned} & Pr(-c \leq T \leq c) = 0.95 \\ \Rightarrow \quad & Pr(-c \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq c) = 0.95 \\ \Rightarrow \quad & Pr(\bar{X} - c\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + c\frac{\sigma}{\sqrt{n}}) = 0.95 \end{aligned} \tag{34}$$

This is a theoretical confidence interval. For an experiment, there are no longer probabilistic concepts attached to the problem and the confidence interval inferred from the experiment is $\left[\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}}\right]$, with $c\frac{\sigma}{\sqrt{n}}$ being called the margin. For $\alpha = 0.05$, $c = 1.96$ or $c = 2$ is a good approximation.

Question 1: A sample of size $n = 100$ produced the sample mean of $\bar{X} = 16$. Assuming the population standard deviation $\sigma = 3$, compute a 95% confidence interval for the population mean μ.

Solution 1: the 95% confidence interval for μ is

$$16 \pm 1.96 \times \frac{3}{\sqrt{100}} = [15.412, 16.588] \tag{35}$$

Question 2: Assuming the population standard deviation $\sigma = 3$, how large should a sample be to estimate the population mean μ with a margin of error not exceeding 0.5 for a 95% confidence level?

Solution 2: We have the margin Δ

$$\begin{aligned} \Delta &\leq c\frac{\sigma}{\sqrt{n}} \\ \Rightarrow n &\geq \left(c\frac{\sigma}{\Delta}\right)^2 \\ \Rightarrow n &\geq \left(1.96\frac{3}{0.5}\right)^2 = 138.3 \end{aligned} \tag{36}$$

4 Bias/unbiased estimators

The bias of an estimator is the difference the estimator's expected value and the true value of the parameter being estimated. We assume $\{X_1, X_2, \ldots, X_n\}$ to be independent and identically distributed (i.i.d.) random variables with expected value μ and variance σ^2. Let's consider the estimator of the mean $M = \frac{1}{n}\sum_{i=1}^{n} X_i$:

Question 1: Compute $E[M]$

Solution 1:

$$\begin{aligned} E[M] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned} \tag{37}$$

Question 2: Is M a bias estimator (i.e is $E[M] - \mu \neq 0$)?

Solution 2: $E[M] = \mu$ therefore it is unbiased

Question 3: Let's consider now this estimator of the variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$. Compute $E[S^2]$. This formula can be helpful

$$Var[M] = Var[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \tag{38}$$

Solution 3:

$$\begin{aligned} E[S^2] &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - M)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu + \mu - M)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2 + 2(X_i - \mu)(\mu - M) + (\mu - M)^2] \\ &= \sigma^2 - E[(\mu - M)^2] \\ &= \sigma^2 - Var[M] \end{aligned} \tag{39}$$

We have

$$\begin{aligned} Var[M] &= Var[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{40}$$

Therefore

$$\begin{aligned} E[S^2] &= \sigma^2 - Var[M] \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2 \end{aligned} \tag{41}$$

Question 4: Is S^2 biased? What estimator of the variance would not be biased?

Solution 4: S^2 is biased but $\frac{n}{n-1}S^2$ is not.

5 Information theory

The Shannon entropy has introduced in 1948 by Claude Shannon and it characterize the average amount of information contained in data produced probabilistically. If X is a discrete random variable then its entropy is defined as

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \tag{42}$$

Question 1: What average information is contained in a random variable such that $P(X = x) = 1$?

Solution 1: $H(X) = -1 \log_2 1 = 0$

Question 2: Let's define a random variable $X = \{1, 2, \ldots, n\}$ such that $P(X = k) = 1/n$ for all $k \in \{1, 2, \ldots, n\}$. Compute $H(X)$

Solution 2:

$$\begin{aligned} H(X) &= -\sum_{x \in X} P(x) \log_2 P(x) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} \\ &= \log_2 n \end{aligned} \tag{43}$$

Question 3: Should we increase or decrease n to increase the average amount of information?

Solution 3: Increase!

We define the mutual information as a measure of probabilistic dependency between 2 random variables X and Y

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right) \tag{44}$$

Question 4: If 2 random variables X and Y are independent we have $P(x,y) = P(x)P(y)$. Compute $I(X;Y)$ if X and Y are independent.

Solution 4:

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right) \\ &= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 1 \\ &= 0 \end{aligned} \tag{45}$$

Question 5: Compute $I(X;Y)$ if $X = Y$ (i.e completely dependent)

Solution 5: We have $P(x) = P(y)$ and $P(x,y) = P(x)$

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right) \\ &= \sum_{x \in X} P(x) \log_2 \left(\frac{P(x)}{P(x)P(x)} \right) \\ &= -\sum_{x \in X} P(x) \log_2 P(x) \\ &= H(X) \end{aligned} \tag{46}$$