Ma 3b Practical – Final Review

March 14, 2025

Please note that while I have proofread these notes, there may still be typos present. These notes are primarily intended for your practice, and there might be occasional typos in the formulas. For complete accuracy, you should refer to the textbook. I cannot be held responsible for any typos that might appear in this materials.

1 Axioms of Probability

1.1 Properties of operators on Events

1. Commutativity: $E \cup F = F \cup E$ and $E \cap F = F \cap E$
2. Associativity: $(E \cup F) \cup G = E \cup (F \cup G)$ and $(E \cap F) \cap G = E \cup (F \cap G)$
3. Distributivity: $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$ and $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$
4. De Morgan's laws: If Γ is a finite or countable index set, then

$$\left(\cup_{\gamma\in\Gamma} E_\gamma\right)^c = \cap_{\gamma\in\Gamma} E_\gamma^c \text{ and } \left(\cap_{\gamma\in\Gamma} E_\gamma\right)^c = \cup_{\gamma\in\Gamma} E_\gamma^c$$

1.2 Axioms of probability

1. From Definition, we have:
   - $P(E) \geqslant 0$ for all $E \in \mathcal{F}$
   - $P(\Omega) = 1$
   - For all mutually disjoint events $\{E_i\}_{i=1}^{\infty}$ (i.e. $E_i \cap E_j = \emptyset$ for all $i \neq j$)

$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

2. Direct sequences

- $P(E^c) = 1 - P(E)$
- $P(\emptyset) = 0$
- $E \subset F$ then $P(E) \leqslant P(F)$
- Sub-additivity: (Note that here we do not require them to be disjoint)

$$P(\cup_{i=1}^{\infty} E_i) \leqslant \sum_{i=1}^{\infty} P(E_i)$$

3. Inclusion-exclusion principle
   - For any event E and F

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

   - In general,

$$P(E_1 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} \cap E_{i_2}) + \cdots + (-1)^{n+1} P(E_1 \cap \cdots \cap E_n)$$

1.3 Exercises

Exercise 1. Find the simplest expression for the following events:

1. $(E \cup F) \cap (E \cup F^c)$
2. $(E \cup F) \cap (E^c \cup F) \cap (E \cup F^c)$
3. $(E \cup F) \cap (F \cup G)$

Solution.

1. $(E \cup F) \cap (E \cup F^c) = E \cup (F \cap F^c) = E \cup \emptyset = E$
2. $(E \cup F) \cap (E^c \cup F) \cap (E \cup F^c) = [(E \cap E^c) \cup F] \cap (E \cup F^c) = [\emptyset \cup F] \cap (E \cup F^c)$
   $= F \cap (E \cup F^c) = (F \cap E) \cup (F \cap F^c) = (F \cap E) \cup \emptyset = E \cap F$
3. $(E \cup F) \cap (F \cup G) = (F \cup E) \cap (F \cup G) = F \cup (E \cap G)$

2 Bayes' Theorem

2.1 Formula

1. Conditional probability: for $P(F) > 0$,
$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}.$$
2. Law of total probability: $P(A) = P(A \cap B) + P(A \cap B^c)$
3. Bayes' theorem:
$$P(A \mid B) = \frac{P(A \cap B)}{P([B)} = \frac{P(B \mid A)P(A)}{\mathbb{P}(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$
4. Independent events: E and F are independent if
$$P(E \cap F) = P(E) \cdot P(F).$$
Or equivalently, $P(E \mid F) = P(E)$. (conditional on F does not give you any new information about E)

2.2 Exercises

Exercise 2. 98% of babies survive childbirth. However, 15% of births require a C-section, and when it does, 96% of babies survive. If a randomly chosen pregnant woman does not have a C-section, what is the probability that her baby will survive?

Solution. For a given childbirth, let
$$E = \{\text{the baby survives}\}$$
$$F = \{\text{there was a caesarean}\}$$
so the information of the problem tells us that
$$P(E) = 0.98 \text{ and } P(F) = 0.15 \text{ and } P(E \mid F) = 0.96$$
By the formula of total probability,
$$P(E) = P(E \mid F^c)P(F^c) + P(E \mid F)P(F)$$
Therefore, we deduce that
$$P(E \mid F^c) = \frac{0.98 - 0.96 * 0.15}{1 - 0.15} = 0.9835.$$

3 Random Variables and Common Distribution

3.1 Discrete random variables

1. Probability mass function (pmf) of X: $p(k) = P(X = K) =$
2. Cumulative distribution function (cdf) of X:
   - $F_X(b) = P(X \leqslant b)$, $b \in \mathbb{R}$
   - 3 properties: (Also for the continuous random variables)
     (a) $\lim_{b\to\infty} F(b) = 1$
     (b) $\lim_{b\to-\infty} F(b) = 0$
     (c) F is right-continuous
   - Summing pmf we get cdf and the jump of cdf corresponds at $X = k$ to pmf at k.
3. Expectation and Variance (Proposition (c) to (e) also holds for continuous r.v.)
   (a) for a discrete r.v. X, the expectation of X is
   $$\mathbb{E}[X] = \sum_{x:p(x)>0} x \cdot p(x)$$
   (b) Proposition: for $g : \mathbb{R} \to \mathbb{R}$ and a r.v. X, consider the random variable $g(X) = g \circ X$. Then,
   $$\mathbb{E}[g(X)] = \sum_{x} g(X)p(x)$$
   (c) Proposition (linearity of expectation)
   $$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$
   the variance of X is defined as
   $$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$
   (d) Proposition:
   $$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$
   (e) Proposition:
   $$\mathrm{Var}(aX + b) = a^2 \cdot \mathrm{Var}(X)$$
   (f) Proposition: If X,Y are independent then
   $$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[g(Y)]\mathbb{E}[g(Y)]$$

3.2 Continuous random variables

1. Probability density function (pdf) $f_X(x)$ of X is a negative function s.t. for all $B \subset \mathbb{R}$
$$P(X \in B) = \int_B f_X(x)dx$$
2. Cumulative distribution function (cdf) of X:
$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$
3. Integrating pdf over the interval $(-\infty, x]$ we get cdf and differentiating cdf we get pdf.
4. Expectation: $\mathbb{E}[x] = \int_{\mathbb{R}} x f_X(x)dx$
5. Proposition: for $g : \mathbb{R} \to \mathbb{R}$ and a r.v. X, consider the random variable $g(X) = g \circ X$. Then,
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)dx$$
6. Variance: $\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$

3.3 Common Distribution Table

Distribution	PMF P(X=k)/ Density f(x)	E	Var
Poisson distribution Pois(λ) (discrete)	$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Geometric distribution Geo(p) (discrete)	$P(X = k) = (1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Bernoulli distribution Bern(p)(discrete)	$P(X = k) = p^k(1-p)^{n-k}(k = 0, 1)$	p	p(1-p)
Binomial distribution Bin(n,p)(discrete)	$P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$	np	np(1-p)
Uniform distribution Unif (a,b) (continuous)	$f(x) = \frac{1}{b-a}$, $x \in [a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal distribution $N(\mu, \sigma^2)$ (continuous)	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ	σ^2
Exponential $Exp(\lambda)$ (continuous)	$f(x) = \lambda e^{-\lambda x}$ if $x \geqslant 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $Gamma(\alpha, \lambda)$ (continuous)	$f(x) = \frac{(\lambda x)^{\alpha-1}\lambda e^{-\lambda x}}{\Gamma(\alpha)}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

Table 1: summary of distributions

Remark: In Gemma distribution, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$

3.4 Exercises

Exercise 3. If X has a cumulative distribution function $F(x) = P(X \leqslant x)$, what is the cumulative distribution function of the random variable e^X?

Solution. Let $Z = e^X$, then $X = \log(Z) = g(Z)$. Note that the function g is increasing,

$$F_Z(z) = P(Z = e^X \leqslant z) = P(g(Z) \leqslant g(z)) = P(X \leqslant \log(z)) = F_X(\log(z)).$$

Exercise 4. Let $X : \Omega ß[0, c]$ be a continuous random variable (not necessarily uniformly distributed !). Show that

$$\mathrm{Var}(X) \leqslant \frac{c^2}{4}$$

Solution. First, we have

$$\mathbb{E}[X^2] = \int_0^c x \cdot xf(x)dx \leqslant \int_0^c c \cdot xf(x)dx = c\mathbb{E}[X]$$

Then define the constant $\alpha = \mathbb{E}[X]/c$, then

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \leqslant c\mathbb{E}[X] - -(\mathbb{E}[X])^2 = c^2\alpha(1-\alpha).$$

Since the function $x \to x(1-x)$ has maximum value $\frac{1}{4}$, so $\mathrm{Var}(X) \leqslant \frac{c^2}{4}$.

4 Basic Statistics Term

4.1 Correlation

1. Covariance of X and Y is defined by:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

2. Proposition: If two r.v. X and Y are independent then $\mathrm{Cov}(X, Y) = 0$

3. Properties:
   - $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
   - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
   - $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$
   - $\mathrm{Cov}(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, Y_j)$
   - $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$

4. Correlation between X and Y is defined by

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

   and $\rho(X, Y) \in [-1, 1]$

4.2 Sample Statistics

From now on, let $X_1, \cdots, X_N$ be i.i.d. random variables from the same distribution F with mean μ and variance σ^2

1. sample mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\mathbb{E}[\overline{X}_n] = \mu$ and $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$
2. sample variance: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, this is an unbiased estimator of σ^2
3. χ^2 distribution: Let $Z_1, \cdots, Z_n$ be i.i.d. $\mathcal{N}(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$
4. Student t distribution: Let $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi_n^2$ and Z is independent of W, then $\frac{Z}{\sqrt{\frac{W}{n}}} \sim t_n$, the student t distribution with degree of freedom n.
5. Normalization of $X \sim \mathcal{N}(\mu, \sigma^2)$: By the property of normal distribution, $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

4.3 Limit distribution

- A random variable obeys a $N(\mu, \sigma^2)$ distribution if its density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
- **Theorem(Weak Law of Large Number)** Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. r.v.s with expectation μ and (finite) variance σ^2, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

  $$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

  for every $\epsilon > 0$.
- **Theorem(Central Limit Theorem)**: Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. r.v.s with expectation μ and (finite) variance σ^2, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, we have the following convergence in law :

  $$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leqslant a\right) = F_Z(a) := \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) dz, \quad \forall a \in \mathbb{R},$$

  where $Z \sim \mathcal{N}(0,1)$. We can also write:

  $$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \leqslant a\right) = F_Z(a), \quad \forall a \in \mathbb{R}.$$

- In general, the Law of Large Number (LLN) says that $\bar{X}_n$ is close to μ, and the Central Limit Theorem (CLT) gives you more information about the distribution: $\bar{X}_n$ converges to $\mathcal{N}(\mu, \sigma^2)$. Or equivalently, LLN gives you the limit shape and CLT gives you the fluctuation.

4.4 Exercises

Exercise 5. 50 numbers are rounded to the nearest integer and then added together. If the rounding errors are uniformly distributed between -0.5 and 0.5 (we assume that rounding errors from different numbers are independent), what is the probability (approximately) that the sum of the 50 rounded numbers differ(in absolute value) from the exact sum by more than 3?

Solution. Let X_i be the rounding error of the i-th number, then $X_i \sim unif(-\frac{1}{2}, \frac{1}{2})$. This implies that $E(X_i) = 0$ and $Var(X_i) = \frac{1}{12}$. And X_i are i.i.d. distributed so we can apply the CLT, then we have approximately

$$\sum_{i=1}^{50} X_i \sim \mathcal{N}(0, \frac{50}{12})$$

Therefore, we have that

$$P(|\sum_{i=1}^{50} X_i| > 3) = P(|Z| > \frac{3}{\sqrt{50/12}}) = P(|Z| > 1.47) = 2P(Z > 1.47) = 0.1416.$$

5 Estimators and Hypothesis Testing

5.1 Maximum Likelihood Estimator (MLE)

Procedure to get MLE:

1. Write out the the likelihood function $L(\theta) = f(X \mid \theta)$
2. Usually to simplify the derivative form, define $l(\theta) = logL(\theta)$
3. Solve for $\hat{\theta}$ s.t. $\frac{dl}{d\theta}(\hat{\theta}) = 0$
4. To verify it is the maximizer, use the second derivative test ($\frac{d^2l}{d\theta^2}(\hat{\theta}) < 0$)

5.2 Confidence Interval, p-value, and Hypothesis Testing

1. Find the correct statistics:
   From now on, we sample X_i i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, and we want to choose the correct statistics W(only consisting of the known data), and use the distribution of W to do the test.
   For Mean testing:

Test Type	Purpose	Requirement	Statistic
One sample z-test	Test mean of a single normal; standard deviation is known	Normally distributed population; known standard deviation	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
One sample t-test	Test mean of a single normal; standard deviation is unknown	Normally distributed population; unknown std. deviation	$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

For Variance testing: For $X_1, \cdots, X_n$ i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, use the statistics $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

Remark: Please see more on the lecture notes for the difference of the means. i.e. X_i from $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y_j from $\mathcal{N}(\mu_2, \sigma_2^2)$, and to test on $\mu_1 - \mu_2$. The statistics (of choosing Z or t) depend on whether σ_1^2 and σ_2^2 are known or not.

2. Confidence Interval: After finding the correct statistics W for the estimation of a parameter θ (e.g. the mean/variance/...), and $W \sim F$, (e.g. F is student t/normal/χ^2/...). Then from F (usually a table is provided) and a given level α, we could find $F_{\alpha/2}$ and $F_{1-\alpha/2}$.
Then we could rewrite the probability (solving the inequality for θ)

$$P(F_{1-\alpha/2} \leqslant W \leqslant F_{\alpha/2}) = P(a \leqslant \theta \leqslant b)$$

Then the (two-sided) random interval is the confidence interval for θ at the confidence level $1-\alpha$.

example. Here we display how to get the two-sided confidence interval for μ with X_i i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ with know σ^2. As described above, we choose the statistics

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

We obtain the inverse quantiles $Z_{\alpha/2}$ and $Z_{1-\alpha/2} = -Z_{\alpha/2}$ because $\mathcal{N}(0,1)$ is symmetric (here F_β means in the graph of pdf f, you find the point on the x-axis x_0, s.t. the area of f at the right of $X = x_0$ is exactly β). Then we have

$$1-\alpha = P(-Z_{\alpha/2} \leqslant Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant Z_{\alpha/2})$$

$$= P(\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

So this interval $[\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ is what we want.

3. Hypothesis testing and p-value:

(a) H_0: the null hypothesis

(b) H_A: the alternative hypothesis

(c)

	Reject H_0	Fail to Reject H_0 (Also referred to as accepting H_A)
H_0 is true	$P($ reject $H_0 \mid H_0$ is true $) = \alpha$	$P($ do not reject $H_0 \mid H_0$ is true $) = 1-\alpha$
H_A is true	$P($ reject $H_0 \mid H_A$ is true $) = 1-\beta$	$P($ do not reject $H_0 \mid H_A$ is true $) = \beta$

(d) p-value: The computation of p-value is very similar to the Confidence interval. After deciding to use statistics W from F distribution, we compute the realized value v by plugging in the data. Then the (two-sided) p-value is

i. $2 \cdot P(Z > |v|)$ for the Z statistics

ii. $2 \cdot P(T > |v|)$ for the t statistics

iii. $2 \cdot \min\{P(K < v), P(K > v)\}$ for the χ^2_{n-1} statistics

Therefore, the Confidence Interval is about for a given level $1-\alpha$ solving for the inverse quantiles like $Z_{\alpha/2}$, and the p-value is about using the statistics value v to compute its corresponding α (Similarly in hypothesis testing, we compare v with $Z_{\alpha/2}$).

5.3 Exercises

Exercise 6. (Homework 6 Question 3) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample where the X_i are distributed according to a $U(0,\theta)$, i.e. a uniform distribution on the interval $[0,\theta]$ where θ is unknown.

a) find $\mathbb{E}(X)$ if X is $U(0,\theta)$

b) Show that the estimator $\hat{\theta} = 2 \cdot \bar{x}$ is unbiased.

c) Find the maximum likelihood estimator for θ

d) Compare the mean squared error for the two estimators, which one has lower MSE?

Solution.

[a)]As usual, we compute

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \cdot f(x)\, dx = \int_0^\theta x \cdot \frac{d\theta}{\theta} = \frac{\theta}{2}$$

Let $\hat{\theta}_n = 2\bar{x} = \frac{2}{n}(x_1 + \ldots + x_n)$. Then by linearity of expectation,

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}_n) &= \mathbb{E}[\hat{\theta}_n] - \theta \\
&= \mathbb{E}\left[\frac{2}{n}(Xx_1 + \ldots + Xx_n)\right] - \theta \\
&= \frac{2}{n}(n\,\mathbb{E}[X]) - \theta \\
&= \frac{2}{n}\left(\frac{n\theta}{2}\right) - \theta \\
&= 0
\end{aligned}$$

So $\hat{\theta}_n$ is unbiased. For convenience, we let the density function of $U(0,\theta)$ be given by

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & 0 \leqslant x \leqslant \theta \\ 0 & \text{else} \end{cases}$$

Therefore since $x_1, \ldots, x_n \geqslant 0$ always, the likelihood (θ) is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i \mid \theta) \\ &= \prod_{i=1}^{n} \begin{cases} \frac{1}{\theta} & 0 \leqslant x_i \leqslant \theta \\ 0 & \text{else} \end{cases} \\ &= \begin{cases} \frac{1}{\theta^n} & \theta \geqslant \max\{x_1, \ldots, x_n\} \\ 0 & \text{else} \end{cases} \end{aligned}$$

and the maximum likelihood estimator $\hat{\theta}_n(x_1, \ldots, x_n)$ is given by $\boxed{\max\{x_1, ..., x_n\}.}$
We compare $MSE(\hat{\theta})$ in the two cases below.

4. • *The estimator* $\hat{\theta} = 2\bar{x}$.
By part (b) we get $\mathbb{E}[\hat{\theta}] = \theta$. We also have

$$\begin{aligned} \mathbb{E}[\hat{\theta}^2] &= \frac{4}{n^2}\mathbb{E}[(X_1 + \ldots + X_n)^2] \\ &= \frac{4}{n^2}\Big(\sum_{i=j} \underbrace{\mathbb{E}[X_i X_j]}_{\mathbb{E}[X^2]} + \sum_{i \neq j} \underbrace{\mathbb{E}[X_i X_j]}_{\mathbb{E}[X]^2}\Big) \\ &= \frac{4}{n^2}\Big(n \underbrace{\mathbb{E}[X^2]}_{\int_0^\theta x^2 \frac{d\theta}{\theta}} + n(n-1)\,\mathbb{E}[X]^2\Big) \\ &= \frac{4}{n^2}\left(n \cdot \frac{\theta^2}{3} + (n^2 - n)\,\frac{\theta^2}{4}\right) \\ &= \frac{\theta^2}{3n} + \theta^2 \end{aligned}$$

Therefore

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 \\ &= \left(\frac{\theta^2}{3n} + \theta^2\right) - 2\theta\,(\theta) + \theta^2 \\ &\boxed{= \frac{\theta^2}{3n}} \end{aligned}$$

- *The maximum likelihood estimator.*
  Recall from part (c) that $\hat{\theta} = \max\{x_1, \ldots, x_n\}$. For $y \in [0, \theta]$, the CDF of $\hat{\theta}$ is hence given by

$$P[\hat{\theta} \leqslant y] = P[X_1, \ldots, X_n \leqslant y] = \left(\frac{y}{\theta}\right)^n$$

  Then the density of $\hat{\theta}$ is given by

$$f_{\hat{\theta}}(y) = \frac{d}{dy}\frac{y^n}{\theta^n} = \frac{ny^{n-1}}{\theta^n}$$

  Finally,

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \int_0^\theta (y - \theta)^2 \, \frac{ny^{n-1}}{\theta^n} \, dy = \boxed{\frac{2\theta^2}{n^2 + 3n + 2}}$$

Assume $\theta > 0$. Comparing the two MSE's above, we find that

$$\begin{cases} MSE(\hat{\theta}_{2\overline{x}}) = MSE(\hat{\theta}_{MLE}) & \text{when } n = 1, 2, \\ MSE(\hat{\theta}_{2\overline{x}}) > MSE(\hat{\theta}_{MLE}) & \text{when } n \geqslant 3. \end{cases}$$

In short, the maximum likelihood estimator has the lower MSE.

Exercise 7. In a certain chemical process, it is extremely important that a certain solution that will be used as a reagent (or reactant) has a pH of exactly 8.20. A method for determining the pH of solutions of this type is known to give measurements with a $\mathcal{N}(\mu, \sigma^2 = 0.0004)$ distribution, where μ represents the current pH of the solution tested. Suppose we take 10 independent pH measurements of a certain solution and observe :

$$8.18, 8.16, 8.17, 8.22, 8.19, 8.17, 8.15, 8.21, 8.16, 8.18.$$

If $H_0 : \mu = \mu_0 = 8.20$, what conclusion can we draw at the level of significance
(a) $\alpha = 0.10$? (b) $\alpha = 0.05$?

Solution. We have $\sigma = 0.02$ and $n = 10$. Given the data, we can calculate $\overline{x}_n = 8.179$. The test statistics (Z-statistics) is thus

$$z = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} = \frac{8.179 - 8.2}{0.02\sqrt{10}} \simeq -3.32.$$

(a) we have $|z| > Z_{\alpha/2} = 1.645$, so we reject H_0 at the significance level $\alpha = 0.10$.
(b) we have $|z| > Z_{\alpha/2} = 1.96$, so we reject H_0 at the significance level $\alpha = 0.05$.

6 Regressions

6.1 Linear Regression

1. X: Response (dependent) variable; Y: Explantory (independent) variable
2. Sample linear correlation coefficient $R_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \in [-1,1]$. The larger $|R_{X,Y}|$ is, the more the scatter plot looks like a straight line.
3. Model $\hat{Y}_i = \hat{a} + \hat{b}X_i$, and the difference $\epsilon_i = Y_i - \hat{Y}_i$ is the residual/error.
4. sum of squared residuals: $SS_R = \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2$ measures the "closeness" of regression line to the data
5. Using the Method of least squares. i.e. solving that $\frac{\partial SS_R}{\partial \hat{a}} = 0$ and $\frac{\partial SS_R}{\partial \hat{b}} = 0$, we obtain that
   - $\hat{a} = \overline{Y} - \hat{b}\,\overline{X}$
   - $\hat{b} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n} X_iY_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$

   where
   $$S_{XY} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n} X_iY_i - n\overline{X}\cdot\overline{Y}$$
   $$S_{XX} = \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$
6. Coefficient of determination:
   $$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}} \in [0,1]$$
   The larger R^2 is, the better is the model. Here we could also interpret as:
   $S_{YY} = SS_R + (S_{YY} - SS_R)$, where SS_R is explained by the error terms $\epsilon_1, \cdots, \epsilon_n$ and $(S_{YY} - SS_R)$ is explained by the input data $x_1, \cdots, x_n$.
   By an exercise, we can show that $R^2_{X,Y} = R^2$.
7. Distribution of the estimators: Normally we assume that $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and we treat the data $x_1, \cdots, x_n$ are constants. We also believe that there exist certain unknown a and b such that $Y = a + bx + \epsilon$. Then we have:
   (a) Can treat $Y_i = a + bx_i + \epsilon_i \sim \mathcal{N}(a + bx_i, \sigma^2)$ as random variables.
   (b) σ^2 is unknown with unbiased estimator $S^2 = \frac{SS_R}{n-2}$ (lose 2 degrees of freedom due to estimating a and b) (you can show that $\frac{SS_R}{\sigma^2} \sim \chi^2_{n-1}$
   (c) $\hat{b} = \frac{\sum_{i=1}^{n} X_iY_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \sim \mathcal{N}(b, \frac{\sigma^2}{S_{xx}})$ is an unbiased estimator for b.

(d) $\hat{a} = \overline{Y} - \hat{b}\,\overline{x} \sim \mathcal{N}(a, \frac{\sigma^2 \sum_{i=1} x_i^2}{nS_{xx}})$ is an unbiased estimator for a.

8. Given the above distribution, we could do the inference on all the parameters. Indeed,

   (a) for b: Use statistic $\frac{\hat{b}-b}{\sqrt{\sigma^2/S_{xx}}\sim\mathcal{N}(0,1)}$

   (b) for a: Use statistic $\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_{i=1}^n x_i^2}}(\hat{a} - a) \sim t_{n-2}$

6.2 Nonlinear Regression: Logistic Regression

In general, there are a lot more models to consider: For example, multiple linear regression (with more than 1 input variable), and some nonlinear models for example, logistic regression, exponential regression, poisson regression, and generalized linear models.

Now we want to use Y_i to help us categorize items, so we make that Y_i can only take values in $\{0, 1\}$, for example, when you want to estimate married women in the labor force, and $Y_i = 0$ means unmarried and $Y_i = 1$ married.

Hence, since $Y_i \in \{0, 1\}$, it is a Bernoulli r.v. with parameter $\pi_i = P(Y_i = 1)$ (Here again we either treat x_i as determined constants or we could let X_i to be random and consider $\pi_i = P(Y_i = 1 \mid X_i = x_i)$ as conditional probability).

But when $Y_i = 1$, $\epsilon_i = 1 - a - bX_i$; and when $Y_i = 0$, $\epsilon_i = -a - bX_i$, so ϵ_i is nonnormal (as an exercise, you can see that the error has nonconstant variance as well). To solve the issue, instead of using the normal distribution, we use logistic distribution for ϵ_L with mean 0 and standard deviation $\sigma = \pi/\sqrt{3}$ with cdf $F_L(\epsilon_L) = \frac{exp(\epsilon_L)}{1+exp(\epsilon_L)}$, we could restate the model as:

Y_i are independent Bernoulli random variables with expected value :

$$\pi_i = E[Y_i] = P(Y_i = 1) = \frac{exp(a + bx_i)}{1 + exp(exp(a + bx_i))}$$

where again we either treat x_i as determined constants or we could let X_i to be random and consider $\pi_i = P(Y_i = 1 \mid X_i = x_i)$ as conditional probability.