Mathematical Statistics - Section 1 - NYU Spring 2019 - Final exam

NAME:

Solutions

- In order to get full credit, you must show your work / justify your answers. This usually means that you should use words alongside your computations.
- No justification is needed for the True/False questions, but pay close attention to the statement of the questions - some are tricky.

Some (approximate) values you can use, where X is a standard normal random variable

x	$\mathbb{P}(X \geq x)$ where $X \sim \mathcal{N}(0,1)$
0	0.5
0.3	0.39
0.5	0.21
0.9	0.19
1.3	0.1
1.6	0.06
1.7	0.045
2.4	0.01
3.9	5/100 000

Some inequalities you can use:

- Markov's inequality: if X is a random variable with finite first moment, for any $t > 0$ we have
$$\mathbb{P}[|X| \geq t] \leq \frac{1}{t}\mathbb{E}[|X|].$$
- Bienaymé-Tchebychev's inequality: if X is a random variable with finite second moment, for any $t > 0$ we have
$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{1}{t^2}\mathrm{Var}[X].$$
- Hölder's inequality (weak form): if $X_1, \ldots, X_m$ are i.i.d. Bernoulli random variables with parameter τ, we have
$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \tau\right| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 m\right).$$

Solution : See HW5

MLE and Fisher information (from HW5) In this exercise, we will denote by M the constant

$$M := \int_{-\infty}^{+\infty} e^{-x^4}\,dx,$$

which we will not try to compute.

We consider the family of probability distribution functions $x \mapsto f(x;\gamma)$ on $(-\infty,\infty)$ defined by

$$f(x;\gamma) := \frac{1}{M}\gamma e^{-\gamma^4 x^4},$$

where γ is some real number.

1. Let $X_1,\ldots,X_n$ be an observation, with true parameter $\gamma_\star$. **Compute** the max-likelihood estimator of $\gamma_\star$

2. **Show** that the Fisher information of this statistical model can be written as

$$I(\gamma) = \frac{A}{\gamma^2},$$

for some constant A that we will not try to compute.

Testing the parameter of an exponential distribution (adapted from HW7)
Let $X_1, \ldots, X_n$ be identically distributed, independent random variables.

We assume that they are distributed according to the exponential distribution on $(0, +\infty)$ given by

$$x \mapsto f(x;\theta) := \theta e^{-\theta x},$$

for a certain value of the parameter $\theta > 0$.

We believe that $\theta = 1$, this is our null hypothesis H_0.

1. **Build** a test statistic T, and an appropriate rejection region R such that
   - The probability of making a type I error is bounded by 10%
   - If $\theta \neq 1$, the probability that T is in R tends to 1 as $n \to \infty$.

Let X have distribution $x \mapsto f(x;\theta)$.

and $\mathrm{Var}(X) = \frac{1}{\theta^2}$

We know that $\mathbb{E}[X] = \frac{1}{\theta}$ (computation done many times). If $X_1, .., X_n$ is an observation, we thus have

$\mathbb{E}[X^2] = \frac{2}{\theta^2}$

$$\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{\theta}\right)\frac{\sqrt{n}}{\theta} \xrightarrow[n\to\infty]{\text{distn}} \mathcal{N}(0,1) \text{ by CLT.}$$

Take $T = \left(\frac{1}{n}\sum_{i=1}^n X_i - 1\right)\cdot\sqrt{n}$, and $R = \{|x| \geq c\}$ for some c to be chosen later.

* Type I error $\quad \mathbb{P}(|T| \geq c) = \mathbb{P}(|\mathcal{N}(0,1)| \geq c)$ (at least as $n \to +\infty$)

Taking $c = 1.7$ we see from the table of values that $\mathbb{P}(|\mathcal{N}(0,1)| \geq 1.7) \leq 0.1$ and thus $\mathbb{P}(\text{Type I error}) \leq 10\%$.

* If $\theta \neq 1$, we have $T \approx \left(\frac{1}{\theta} - 1\right)\cdot\sqrt{n}$, again by CLT,

so $|T| \xrightarrow[n\to\infty]{} +\infty$ in probability.

and $\mathbb{P}(|T| \geq c) \xrightarrow[n\to\infty]{} 1$.

This can be made more rigorous.

2. If we remove either one of these two properties, why is the question much simpler?

if we only want to control Type I $\to$ take $R = \emptyset$

Type II $\to$ take $R = \mathbb{R}$.

CLT and applications

1. Let $X_1, \dots, X_n$ be i.i.d. random variables such that $\mathbb{E}[X_1^2]$ is finite. **Compute**

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i)$ because the X_i's are $\perp\!\!\!\perp$

Since they are identically distributed, we get

$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\operatorname{Var}(X_1)$.

2. **State** the central limit theorem.

See notes.

3. Application: Two candidates are running for election (the one who gets the most votes is elected) in a country with a billion voters. After counting a million votes, the results are as follows:

$$A : 532\ 000, B : 468\ 000.$$

Let p be the proportion of voters for A.

Who do you think is going to win? Justify your answer with a quantitative argument, e.g. a confidence interval, a p-value for a certain test...

$\hat{p}$ the natural estimator

→ I think A is going to win.

The standard deviation of $\hat{p}$ is bounded by $\frac{\sqrt{\frac{1}{4}}}{\sqrt{n}} = \frac{1}{2 \cdot 1000}$

An asymptotic confidence interval is thus given by:

$p \in \left[0.532 - \frac{3.9}{2000}, 0.532 + \frac{3.9}{2000}\right]$

with confidence $1 - \frac{1}{10000}$, as given by the table of values.

This interval does not contain $\frac{1}{2}$.

We can say that A is going to win with at least $1 - \frac{1}{10000}$ confidence (in fact, more than that).

Statistics Below is an excerpt of a data table from the "Demographic and Health Surveys", concerning the mean height (in centimeters) for adult women in various countries. SD stands for (empirical) "standard deviation". "Percent Urban" is the percentage of the women in the sample who live in a city.

Country	Sample size	Empirical mean height	SD height	Percent Urban
Azerbaidjan	5,412	158.4	5.9	52.9
Benin	11,015	159.3	6.5	40.3
Colombia	22,947	155.0	6.2	76.4

1. Do you think that the average height among all women in Azerbaidjan is greater than the average height among all women in Colombia? **Justify** your answer with a quantitative argument, e.g. a confidence interval, a p-value for a certain test...

~~I don't think so. Take H_0 = "average height women in A > average height in B", so the difference $h_A - h_B \geq 0$. An estimator for $h_A - h_B$ is given by $158.4 - 159.3$ and the standard deviation is $\sqrt{\frac{5.9^2}{5412} + 6.}$~~

* of the estimator

Yes. The difference $h_A - h_C$ between average heights can be estimated by $158.4 - 155.0 = 3.4$. The standard deviation* is $sd = \sqrt{\frac{5.9^2}{5412} + \frac{6.2^2}{22947}}$

$$sd \approx \sqrt{\frac{25}{5000} + \frac{36}{20000}} \approx \sqrt{\frac{1}{200} + \frac{1}{500}} \approx \sqrt{\frac{3}{500}} \approx \sqrt{\frac{4}{400}} \approx \frac{1}{10}$$ (rough estimate)

So we are 34 standard deviations away from equality...

A confidence interval is $h_A - h_C \in \left[3.4 - \frac{3.9}{10}, 3.4 + \frac{3.9}{10}\right]$ with $1 - \frac{1}{10000}$ confidence. Does not contain 0!

2. If we gather the data from these three samples, will the (empirical) standard deviation of heights be $\frac{5.9+6.5+6.2}{3} = 6.2$?

No, that's not how it works. Take a sample of size 1, and another one of size one million with a big standard deviation. The standard deviation of the gathered sample will still be big, not half the previous one. The exact computation is not totally obvious, but there is no reason to get 6.2.

3. Can we say that living in cities causes people to grow taller?

No. It would require a much more careful analysis anyway, but here the correlation between "urban" and "average height" is negative! It's impossible to tell just from this data.

Solution: See HW5

MLE and Fisher information (from HW5) In this exercise, we will denote by M the constant

$$M := \int_{-\infty}^{+\infty} e^{-x^4}\,dx,$$

which we will not try to compute.

We consider the family of probability distribution functions $x \mapsto f(x;\gamma)$ on $(-\infty,\infty)$ defined by

$$f(x;\gamma) := \frac{1}{M}\gamma e^{-\gamma^4 x^4},$$

where γ is some real number.

1. Let $X_1, \ldots, X_n$ be an observation, with true parameter $\gamma_\star$. **Compute** the max-likelihood estimator of $\gamma_\star$

2. **Show** that the Fisher information of this statistical model can be written as

$$I(\gamma) = \frac{A}{\gamma^2},$$

for some constant A that we will not try to compute.

Testing the parameter of an exponential distribution (adapted from HW7)
Let $X_1, \dots, X_n$ be identically distributed, independent random variables.

We assume that they are distributed according to the exponential distribution on $(0, +\infty)$ given by

$$x \mapsto f(x;\theta) := \theta e^{-\theta x},$$

for a certain value of the parameter $\theta > 0$.

We believe that $\theta = 1$, this is our null hypothesis H_0.

1. **Build** a test statistic T, and an appropriate rejection region R such that
   - The probability of making a type I error is bounded by 10%
   - If $\theta \neq 1$, the probability that T is in R tends to 1 as $n \to \infty$.

Let X have distribution $x \mapsto f(x;\theta)$.

We know that $\mathbb{E}[X] = \frac{1}{\theta}$ and $\mathrm{Var}(X) = \frac{1}{\theta^2}$ (Computation done many times). If $X_1, .., X_n$ is an observation, we thus ~~$\mathbb{E}[X^2] = \frac{2}{\theta^2}$~~ have

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\theta}\right)\frac{\sqrt{n}}{\theta} \xrightarrow[n\to+\infty]{\text{distn}} \mathcal{N}(0,1) \text{ by CLT.}$$

Take $T = \left(\frac{1}{n}\sum_{i=1}^{n} X_i - 1\right)\cdot\sqrt{n}$, and $R = \{|x| \geq c\}$ for some c to be chosen later.

* Type I error $\quad \mathbb{P}(|T| \geq c) = \mathbb{P}(|\mathcal{N}(0,1)| \geq c)$ (at least as $n \to +\infty$)

Taking $c = 1.7$ we see from the table of values that $\mathbb{P}(|\mathcal{N}(0,1)| \geq 1.7) \leq 0.1$ and thus $\mathbb{P}(\text{Type I error}) \leq 10\%$

* If $\theta \neq 1$, we have $T \approx \left(\frac{1}{\theta} - 1\right)\cdot\sqrt{n}$, again by CLT, so $|T| \xrightarrow[n\to\infty]{} +\infty$ in probability. and $\mathbb{P}(|T| \geq c) \xrightarrow[n\to+\infty]{} 1$.

This can be made more rigorous.

2. If we remove either one of these two properties, why is the question much simpler?

if we only want to control Type I $\to$ take $R = \emptyset$

Type II $\to$ take $R = \mathbb{R}$.

CLT and applications

1. Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $\mathbb{E}[X_1^2]$ is finite. **Compute**

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i)$ because the X_i's are $\perp\!\!\!\perp$

Since they are identically distributed, we get

$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\operatorname{Var}(X_1)$.

2. **State** the central limit theorem.

See notes.

3. Application: Two candidates are running for election (the one who gets the most votes is elected) in a country with a billion voters. After counting a million votes, the results are as follows:

$$A : 532\ 000, B : 468\ 000.$$

Let p be the proportion of voters for A.

Who do you think is going to win? Justify your answer with a quantitative argument, e.g. a confidence interval, a p-value for a certain test...

$\hat{p}$ the natural estimator

→ I think A is going to win.

The standard deviation of $\hat{p}$ is bounded by $\frac{\sqrt{\frac{1}{4}}}{\sqrt{n}} = \frac{1}{2 \cdot 1000}$

An asymptotic confidence interval is thus given by:

$p \in \left[0.532 - \frac{3.9}{2000}, 0.532 + \frac{3.9}{2000}\right]$

with confidence $1 - \frac{1}{10000}$, as given by the table of values.

This interval does not contain $\frac{1}{2}$.

We can say that A is going to win with at least $1 - \frac{1}{10000}$ confidence (in fact, more than that).

Statistics Below is an excerpt of a data table from the "Demographic and Health Surveys", concerning the mean height (in centimeters) for adult women in various countries. SD stands for (empirical) "standard deviation". "Percent Urban" is the percentage of the women in the sample who live in a city.

Country	Sample size	Empirical mean height	SD height	Percent Urban
Azerbaidjan	5,412	158.4	5.9	52.9
Benin	11,015	159.3	6.5	40.3
Colombia	22,947	155.0	6.2	76.4

1. Do you think that the average height among all women in Azerbaidjan is greater than the average height among all women in Colombia? **Justify** your answer with a quantitative argument, e.g. a confidence interval, a p-value for a certain test...

~~I don't think so. Take H_0 = " average height women in A > average height in B", So the difference $h_A - h_B \geq 0$. An estimator for $h_A - h_B$ is given by 158.4 − 159.3 and the standard deviation is~~ $\cancel{\sqrt{\frac{5.9^2}{5412} + 6.}}$

* of the estimator

Yes. The difference $h_A - h_C$ between average heights can be estimated by $158.4 - 155.0 = 3.4$. The standard deviation* is $sd = \sqrt{\frac{5.9^2}{5412} + \frac{6.2^2}{22947}}$

$$sd \approx \sqrt{\frac{25}{5000} + \frac{36}{20000}} \approx \sqrt{\frac{1}{200} + \frac{1}{500}} \approx \sqrt{\frac{3}{500}} \approx \sqrt{\frac{4}{400}} \approx \frac{1}{10}$$ (rough estimate)

So we are 34 standard deviations away from equality...

A confidence interval is $h_A - h_C \in \left[3.4 - \frac{3.9}{10}, 3.4 + \frac{3.9}{10}\right]$ with $1 - \frac{1}{10000}$ confidence. Does not contain 0!

2. If we gather the data from these three samples, will the (empirical) standard deviation of heights be $\frac{5.9+6.5+6.2}{3} = 6.2$?

No, that's not how it works. Take a sample of size 1, and another one of size one million with a big standard deviation. The standard deviation of the gathered sample will still be big, not half the previous one. The exact computation is not totally obvious, but there is no reason to get 6.2.

3. Can we say that living in cities causes people to grow taller?

No. It would require a much more careful analysis anyway, but here the correlation between "urban" and "average height" is negative! It's impossible to tell just from this data.

True or false? Note: a statement that is *meaningless* should be treated as false.

TF1	TF2	TF3	TF4	TF5	TF6	TF7	TF8	TF9	TF10
F	T	F	F	F	T	F	T	F	F

1. Let g be a quantity and $\hat{g}$ be an estimator of g. If $\hat{g}$ is consistent, then it is unbiased. No.
2. Let $(X_i)_{i\geq 1}$ be random variables, such that $\mathbb{E}[X_i] = 10$ for all i. Then the following convergence holds
$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = 10.$$
Yes, in fact no need to take the limit.
3. Let $(X_i)_{i\geq 1}$ be random variables, such that $\mathbb{E}[X_i] = 10$ for all i. Then the following convergence holds *in probability*
$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = 10.$$
No, we need the X_i's to be $\perp\!\!\!\perp$.
4. Let X, Y be two random variables, with real values. We assume that for every real number t, we have
$$\mathbb{P}[X \geq t, Y \geq 0] = \mathbb{P}[X \geq t]\,\mathbb{P}[Y \geq 0].$$
Then X and Y are independent.
No. If the r.v. are always negative, this is just $0=0$.
5. Let X_0 be some random variable, let N_1, N_2 be two random variables, and let
$$X_1 := X_0 + N_1, \quad X_2 := X_0 + N_2.$$
If N_1 and N_2 are independent, then X_1 and X_2 are independent.
No, take $N_1 = N_2 = 0$.
6. Let N_1, N_2 be two random variables, and let
$$X_1 := \frac{1}{2} + N_1, \quad X_2 := \frac{1}{4} + N_2.$$
If N_1 and N_2 are independent, then X_1 and X_2 are independent.
Yes.
7. The larger the p-value, the more confidence we have when rejecting the null hypothesis.
No, the smaller the p-value...
8. A smaller variance for an unbiased estimator means a narrower confidence interval for the estimated quantity. Yes
9. The Fisher information is the best estimator we have for the empirical mean of a parametric predictor.
↑ Meaningless
10. Let X, Y be two standard normal random variables. Then
$$\underbrace{\mathrm{Var}(10 + X)}_{1} \geq \underbrace{\mathrm{Var}(-5 + 2Y)}_{4}.$$
$1 \geq 4$ No.

Linear regression through the origin . Here is a data set of four data points:

$$(X_1, Y_1) = (1, 3), \quad (X_2, Y_2) = (2, 5), \quad (X_3, Y_3) = (3, 6), \quad (X_4, Y_4) = (5, 12).$$

Find the linear regression through the origin that minimizes the residual sum of squares. In other words, find the coefficient α such that

$$RSS := \sum_{i=1}^{4} (Y_i - \alpha X_i)^2$$

is as small as possible.

Let $\vec{Y} = \begin{pmatrix} 3 \\ 5 \\ 6 \\ 12 \end{pmatrix}$, $\vec{X} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}$ we are looking for the orthogonal proj. of $\vec{Y}$ on the line $\mathbb{R}\vec{X}$, so for α such that

$$(\vec{Y} - \alpha\vec{X}) \cdot \vec{X} = 0$$

$$\alpha\|\vec{X}\|^2 = \vec{Y} \cdot \vec{X} \qquad \alpha = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\|^2}$$

$$\vec{X} \cdot \vec{Y} = 3 + 10 + 18 + 60 = 91$$

$$\|\vec{X}\|^2 = 1 + 4 + 9 + 25 = 39$$

So $\boxed{\alpha = \frac{91}{39} \approx 2.25}$

Can also study the function $\alpha \mapsto RSS(\alpha)$ and take derivative $= 0$ ...

About the empirical cdf (inspired by HW3) Let $x \mapsto f(x)$ be a continuous pdf on $\mathbb{R}$. Let $X_1, \dots, X_n$ be i.i.d. random variables with common distribution f.

For any real numbers t, s with $t < s$, we let $G(t,s)$ be the quantity

$$G(t,s) := \int_t^s f(x)dx.$$

and we define the statistic $\hat{G}_n(t,s)$ as

$$\hat{G}_n(t,s) := \frac{\text{number of data points in } [t,s)}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{t \le X_i < s}.$$

We recall that $\mathbf{1}_{t \le X < s}$ is equal to 1 if $X \in [t,s)$ and to 0 if $X \notin [t,s)$. The variables $\mathbf{1}_{t \le X_i < s}$ $(i = 1 \dots n)$ are thus i.i.d. Bernoulli random variables.

1. For any $t < s$ fixed, **compute** the expectation and the variance of $\hat{G}_n(t,s)$.

Let us observe that $\mathbb{E}[\mathbb{1}_{t \le X < s}] = \mathbb{P}(t \le X < s) = G(t,s)$

So $\mathbb{E}[\hat{G}_n(t,s)] = \frac{n}{n} G(t,s) = \underline{G(t,s)}$ by linearity of $\mathbb{E}$

$\mathrm{Var}[\hat{G}_n(t,s)] = \frac{1}{n} G(t,s)(1 - G(t,s))$ because the X_i's are independent

and $\mathrm{Var}(\mathbb{1}_{t \le X < s}) = G(t,s)(1 - G(t,s))$ [variance of a Bernoulli r.v.]

2. **Prove** that $\hat{G}_n(t,s)$ is an unbiased, consistent estimator of $G(t,s)$. *Hint: you can rely on results that we have proved in class, or you can re-do the proof.*

$G(t,s) = F(s) - F(t)$ for F the cdf defined in class

$\hat{G}_n(t,s) = \hat{F}_n(s) - \hat{F}_n(t)$ for $\hat{F}_n$ the empirical cdf defined in class.

We know from class

$\hat{F}_n(x) \xrightarrow[n \to \infty]{\mathbb{P}} F(x)$ for every x

so

$\hat{G}_n(t,s) \xrightarrow[n \to \infty]{\mathbb{P}} G(t,s)$.

Law of large numbers ✓

For consistency

Unbiasedness was proven in 1.

Bonus question: pick the question you prefer (and answer it):

1. Continuation of the last exercise: **compute** the covariance of $\hat{G}_n(0,1)$ and $\hat{G}_n(0,2)$.
2. Three candidates A, B, C are running for election (the one who gets the most votes is elected) in a country with a million voters. After counting 10 000 votes, the results are as follows:

$$A : 3245, B : 4765, C : 1990.$$

   Who is going to win?
3. Let f be some probability distribution function. We **believe** that it is an exponential distribution of parameter 1. **Explain** how you would design a test for this hypothesis.

3. "Goodness of fit": discretize and use χ^2 (see lecture notes)

1. Observe $\hat{G}_n(0,2) = \hat{G}_n(0,1) + \hat{G}_n(1,2)$ so

$$\mathrm{Cov}(\hat{G}_n(0,1), \hat{G}_n(0,2)) = \mathrm{Var}(\hat{G}_n(0,1)) + \mathrm{Cov}(\hat{G}_n(0,1), \hat{G}_n(1,2)).$$

Write $\mathrm{Cov}(\hat{G}_n(0,1), \hat{G}_n(1,2))$ as $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(\mathbb{1}_{0\leq X_i<1}, \mathbb{1}_{1\leq X_j\leq 2})$

X_i, X_j are $\perp\!\!\!\perp$ for $i \neq j$, it gives $= \frac{1}{n}\mathrm{Cov}(\mathbb{1}_{0\leq X\leq 1}, \mathbb{1}_{1\leq X\leq 2})$

$= \frac{1}{n}\left(-G(0,1)\,G(1,2)\right)$ (from class, or a computation).

So $\mathrm{Cov}(\hat{G}_n(0,1), \hat{G}_n(0,2)) = \frac{1}{n}\left[G(0,1)(1-G(0,1)) - G(0,1)\,G(1,2)\right]$

2. It seems like B is going to win. To "prove" this, we could show that $p_B > p_A$ with good confidence, and $p_B > p_C$ with good confidence.

⚠ Estimators for p_A, p_B, p_C given by $\hat{p}_A, \hat{p}_B, \hat{p}_C$ are not independent...