18.095: The Stein Paradox

Lecturer: Philippe Rigollet Jan. 13, 2016

1. WHAT IS STATISTICS ABOUT?

In a nutshell, statistics is about inferring the properties of a large (potentially infinite population) from the observation of a subset of this population. The idea is that it may be impossible or too costly to query the entire population. Here are some examples.

1.1 Gallup polls

This is perhaps the best known example of statistics at work, especially during an election. The population we are trying to understand are the voters, more specifically, the voters on the day of the election as opinion may change between a poll and an election.

Today, everyone is talking about the Iowa republican caucuses. In this case the population of interest is **registered and voting republicans on the day of the election**. In 2012, this population counted 121,501 voters, which is clearly too big to poll entirely. How are Gallup polls made?

A press release dated Jan. 12, 2016 states

> A new Public Policy Polling surveyed 530 likely GOP caucus goers and found that the GOP presidential front-runner Trump leads with 28%, a few weeks ahead of the first-in-the-nation caucus.

Is 530 a large enough number? How was this number 530 found? We will try to answer these questions.

1.2 Drug discovery

Every year hundreds of new drugs are tested to be released subject to the approval of the US Food and Drug Administration (FDA). Aripiprazole, sold under the brand name Abilify® by Otsuka Pharmaceutical, is an antipsychotic. It is recommended and primarily used in the treatment of schizophrenia and bipolar disorder. Before releasing Abilify® in 2002, Otsuka ran clinical trials to measure the *efficacy* of the drug among other things such as side effects. To be approved by FDA, the clinical trials must demonstrate that the drug is more performant than the benchmark, typically a placebo. Such clinical trials are heavily regulated and typically operate in phases and we focus on the last one before launching the product. In this phase, the most patients possible are enrolled in the study to obtain the best possible accuracy.

Here the population of interest is the set of **all potential future consumers of Abilify®**. This an abstract population, unlike in the previous example.

In the case of bipolar schizophrenic patients, a study was run on 414 patients, measuring some criterion called PANSS (the larger the better). The difference PANSS(placebo)-PANSS(Abilify®) is reported to lie in the following *confidence interval*: $(-14.8, -2.1)$. It is clearly only negative numbers so it is good but how was this interval constructed? Why is it not a single number for PANSS(placebo)-PANSS(Abilify®)? How would have the result changed if we had 4,000 patients in this study?

1.3 Body measurements

In class, we asked 5 students to measure (in cm) the following:

1. span of their right hand
2. length of left forearm and
3. circumference of wrist

What population is it representative of? The students in the classroom on that day? MIT undergrads? MIT students? Americans? People? Mammals? In principle, we can extrapolate as much as we want but the bigger the population the larger sample we have access to.

2. RANDOM VARIABLES

So statistics is about understanding what the population might look like from a few samples. It always goes in pair with *probability*, which instead is trying to look at what *random* samples from a given population will look like. The word `random` is key here: the 530 surveyed persons are chosen at random, the 414 patients are also chosen at random (and assigned drug or placebo at random) and the 5 students were chosen at random. This is very important if we want the sample to be *representative* of the population.

2.1 The binomial distribution

Let us go back first to the poll example. Assume that eventually, 30% of the voters will choose Trump. What will a sample of size 530 look like? Since $530 * 30\% = 159$, we expect about 159 of the respondents in the poll to say Trump. However, in the above poll $148 \simeq 530 * 28\%$ people responded that they would vote Trump. What if we had 160? Is that much less likely than having 160 or 158 Trump supporters in the sample?

Probability theory allows us to *quantify* exactly the probability for each of these outcomes as long as we know the proportion in the overall population. Let us assume for the time being that this number is .3, that is 30% of the republican voters will vote for Trump on republication election day. What is the probability that we will see exactly 159 Trump supporters in our sample of $n = 530$ republican voters? To answer this question, let's go back to basics by changing the size n of our sample. Write N to be the number of Trump supporters.

- Assume first that we have a sample of size $n = 1$. The probability that this person is a Trump supporter is $\mathbb{P}(N) = \mathbb{P}(T) = 0.3$
- Assume that $n = 2$ and our sample as 2 voters. Let's call them Bernie and Hillary. What is the probability that one of these two is a Trump supporters? This can happen if Bernie is a Trump supporter and Hillary is not (let's write this as $T\bar{T}$) or vice versa if Hillary is a Trump supporter and Bernie is not ($\bar{T}T$). From the randomness of our sample, Hillary's vote is *independent* of that of Bernie's. So that $\mathbb{P}(T\bar{T}) = \mathbb{P}(T)\mathbb{P}(\bar{T})$, therefore[1]

$$\mathbb{P}(N = 1) = \mathbb{P}(T\bar{T} \text{ or } \bar{T}T) = 2\mathbb{P}(T)\mathbb{P}(\bar{T}) = 2\mathbb{P}(T)[1 - \mathbb{P}(T)] = 2 \cdot 0.3 \cdot 0.7 = 0.42$$

[1] here we used the two basic laws of probability:

1. $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$, that is A and B are disjoint ($T\bar{T}$ and $T\bar{T}$ are disjoint) and

- Assume now that $n = 3$ and we want to compute $\mathbb{P}(N = 1)$. One voter supporting Trump can happen in one of three ways: $T\bar{T}\bar{T}$, $\bar{T}T\bar{T}$ or $\bar{T}\bar{T}T$. Each of these disjoint events have probability $0.3(0.7)^2 = 0.147$ so that $\mathbb{P}(N = 1) = 3 \cdot 0.147 = 0.441$.

More generally if we have n voters, what is $\mathbb{P}(N = k)$, $k = 0, \ldots, n$? We need basic combinatorics first: we need to count sequences of Ts and $\bar{T}$s of length n contain exactly k Ts. The answer is the well known quantity

$$\binom{n}{k} = \frac{n!}{k! * (n-k)!} = \frac{n \cdot (n-1) \cdots 2 \cdot 1}{k \cdot (k-1) \cdots 2 \cdot 1 * (n-k) \cdot (n-k-1) \cdots 2 \cdot 1}$$
$$= \frac{n \cdot (n-1) \cdots (n-k+2) \cdot (n-k+1)}{k \cdot (k-1) \cdots 2 \cdot 1}$$

To convince yourself that this is true, note that we have to place k T's in n possible slots. For the first one, there are all n slots. For the second, there are $n-1$ remaining slots, etc. This gives the numerator. To explain the presence of $k!$ in the denominator, note that we have counted some sequences several times. In the case $N = 3, k = 2$, we can easily see why by looking at these two scenarios:

1. The first T goes to position 1, the second goes to position 2
2. The first T goes to position 2, the second goes to position 1

Clearly both scenarios form the sequence $TT\bar{T}$, which was counted twice. for a general k, how many times was it counted? The answer is: once for each possible order of the k Ts. There are $k! = k(k-1) \cdots 2 \cdot 1$ such orders. Indeed, there are k choices for the first one, $k-1$ choices for the second one etc.

So we have determined that there are $\binom{n}{k}$ sequences with exactly k Ts. What is the probability of such a sequence? All have the same probability so let us take the simplest one to compute: $\underbrace{TT\cdots T}_{k}\underbrace{\bar{T}\bar{\cdots}\bar{T}}_{n-k}$. We have, by independence

$$\mathbb{P}(TT\cdots T\bar{T}\bar{\cdots}\bar{T}) = \underbrace{\mathbb{P}(T)\mathbb{P}(T)\cdots\mathbb{P}(T)}_{k}\underbrace{\mathbb{P}(\bar{T})\mathbb{P}(\bar{T})\cdots\mathbb{P}(\bar{T})}_{n-k} = \mathbb{P}(T)^k(1-p(T))^{n-k}\,.$$

Therefore, for each n and k, we have a formula

$$\mathbb{P}(N = k) = \binom{n}{k} 0.3^k 0.7^{n-k}$$

So N is a random number between 0 and n and we know the probability for each value that it can take. When it satisfies this formula, we say that N has a *binomial distribution* with parameters n and $p = 0.3$. We can also say (less precisely) that N is a binomial random variable. For a general $p \in (0, 1)$, not necessarily equal to 0.3 and a general n, we have

$$\mathbb{P}(N = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In this case, we write $N \sim \mathsf{Bin}(n, p)$.

2. if A and B are independent $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

Together these imply that $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$

In our numerical example, $n = 530$ so

$$\mathbb{P}(N = 159) = \binom{530}{159} 0.3^{159} 0.7^{371} = 0.038\,,$$

and $\mathbb{P}(N = 158) = 0.037$ for example. The probabilities $\mathbb{P}(N = k)$ for $k = 0, \ldots, 530$ are plotted in Figure 1. Note that they sum up to 1.

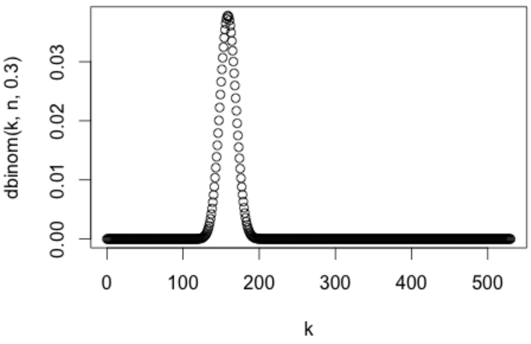

Figure 1: Binomial probabilities for parameters $n = 530$ and $p = 0.3$

2.2 The Bernoulli distribution

There exists a nice and convenient way to represent a binomial N. Recall that in our example N is the number of voters among $n = 530$ randomly chosen that intend to vote for Trump. For each voter $i = 1, \ldots, n$, let X_i be a random variable takes only one of two values:

$$X_i = \begin{cases} 1 & \text{if voter } i \text{ intends to vote for Trump} \\ 0 & \text{otherwise} \end{cases}$$

It is not hard to see that we can represent

$$N = \sum_{i=1}^{n} X_i \tag{2.1}$$

(indeed only the ones that vote for Trump are summed together). Note that for each voter i, $\mathbb{P}(X_i = 1) = 0.3$ and therefore $\mathbb{P}(X_i = 0) = 0.7$. Such a random variable X_i, that takes only two values is said to have *Bernoulli*[2] *distribution* with parameter 0.3. More generally a random variable $X \in \{0, 1\}$ such that $X = 1$ with probability p and $X = 0$ with probability $1 - p$ for some $p \in [0, 1]$ is said to have *Bernoulli distribution* with parameter p, or simply that X has Bernoulli distribution and we write $X \sim \mathsf{Bern}(p)$.

Note that if N has a binomial distributions with parameters $n = 1$ and p then N has a Bernoulli distribution with parameter p.

2.3 The Gaussian distribution

When it comes to body measurements, the random variables that we get are not integers but rather real numbers (we neglect rounding effects). Indeed, the length of a forearm could be any real number in a reasonably large interval. Since there is a continuum of values, rather than giving the probability that the measurement is *equal* to some value, we give the probability that it falls in a set $A \subset \mathbb{R}$:

$$\mathbb{P}(X \in A) = \int_A f(x)dx$$

where $f(\cdot)$ is called the density of X. It satisfies $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_R f(x)d(x) = \mathbb{P}(X \in \mathbb{R}) = 1$. Such a random variable is called *continuous* as opposed to the binomial or Bernoulli random variables that are called *discrete.*

[2] named after Swiss scientist Jacob Bernoulli. He derived the first version of the law of large numbers in his work Ars Conjectandi.

The randomness of the body measurements does not arise from first principles like that of the number of Trump supporters. Rather, we make some *modeling assumptions*. This means that we assume that $f \in \mathcal{F}$ is in some class $\mathcal{F}$ of functions. An overwhelmingly popular one is the following class:

$$\mathcal{F} = \Big\{ f_{\mu,\sigma^2} \,:\, f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big), x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0 \Big\}$$

If a random variable X has density f_{μ,σ^2}, we say that it has Gaussian (or Normal) distribution with parameters μ and σ^2 and we write $X \sim \mathcal{N}(\mu, \sigma^2)$. The function $x \mapsto f_{\mu,\sigma^2}(x)$ has a well known bell shaped curve (See figure 2).

A useful property of the Gaussian distribution that can be easily checked using a change of variable is that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

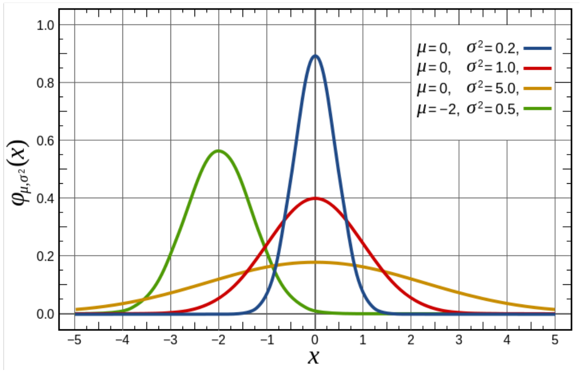

Figure 2: Gaussian densities for various values of μ and σ^2 (source Wikipedia).

So far, we have only talked about random variables but we can also talk about random vectors (even random matrices but this is for another lecture). For example, if we concatenate the three body measurements into on vector of $\mathbb{R}^3$, we get a random vector. Let X_i denote the ith random body measurement for $i = 1, 2, 3$ and assume that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Assume further that the three random variables are independent. Then the vector $X = (X_1, X_2, X_3) \in \mathbb{R}^3$ has *multivariate Gaussian distribution* with density:

$$f(x_1, x_2, x_3) = f_{\mu_1,\sigma_1^2}(x_1) \cdot f_{\mu_2,\sigma_2^2}(x_2) \cdot f_{\mu_3,\sigma_3^2}(x_3)\,.$$

This means that for any $A \subset \mathbb{R}^3$,

$$\mathbb{P}((X_1, X_2, X_3) \in A) = \iiint_A f(x_1, x_2, x_3) dx_1 dx_2 dx_3.$$

For independent random variables, it is always true that the density of the vector that they form is given by the product of their individual (aka marginal) densities.

3. MAXIMUM LIKELIHOOD ESTIMATION

The random variables $N \sim \mathsf{Bin}(530, p)$ or $X \sim \mathcal{N}(\mu, \sigma^2)$ depend on unknown parameters that we would like to estimate from the data that we have collected.

3.1 Binomial distribution

To start, let us go back to our poll. All we know for this problem is that $n = 530$ and that the binomial random variable N was *observed* to be 148 (note that this number is not random). From this we would like to infer the parameter p, the proportion of the population that will eventually vote for Trump. There are many statistical methods to do that, but perhaps the most popular one is called *maximum likelihood estimation*.

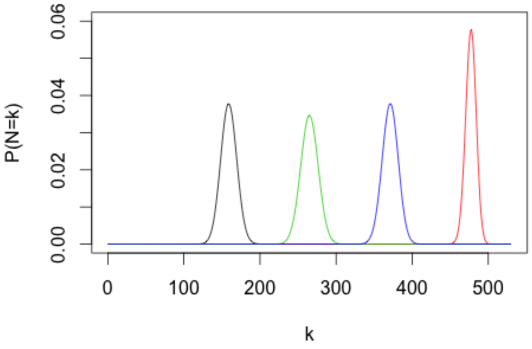

Figure 3: Binomial probabilities for parameters $n = 530$ and $p = 0.3$ (black), $p = 0.5$ (green), $p = 0.7$ (blue) and $p = 0.9$ (red)

The principle behind maximum likelihood estimation is quite simple: choose the parameter which is the most likely to have generated the data at hand. In our poll example, we are trying to find the parameter p which is the most likely to have generated the number 148. In Figure 3, we see that when p varies, the most likely value (the one with largest probability) changes.

The question becomes: what value of $p \in (0,1)$ is such that $\mathbb{P}(N = 148) = \max_{1 \le k \le 148} \mathbb{P}(N = 148)$? We need to study the function

$$p \mapsto \binom{530}{148} p^{148}(1-p)^{382}$$

and find its maximum. We can do this in full generality by studying the function:

$$p \mapsto g(p) = \binom{n}{k} p^k (1-p)^{n-k}$$

While we would need to check that carefully using second derivatives, we can see from Figure 3 that the only extremum is a maximum so we need to find $\hat{p}$ such that the derivative vanishes: $f'(\hat{p}) = 0$.

$$g'(\hat{p}) = \binom{n}{k} \left[k p^{k-1}(1-p)^{n-k} - (n-k)p^k(1-p)^{n-k-1} \right] = 0$$

this is equivalent to

$$\hat{p} = \frac{k}{n}\,.$$

In other words, the maximum likelihood estimator is simply the average number of Trump supporters within the sample. This is reassuring to get such a simple estimator in the binomial case but there exist models where the maximum likelihood estimator may be much more complicated.

Applying this to our numerical data, we get $\hat{p} = 148/530 \simeq 0.28$.

3.2 Gaussian distribution

For the Gaussian distribution, there are 2 unknown parameters μ and σ^2 for each measurement. To make things simple, assume that $\sigma^2 = 1$ and our goal is to estimate μ. Consider the span of the hand for example. Our observations consist of a random vector $X = (X^{(1)}, \dots, X^{(5)}) \in \mathbb{R}^5$ where the coordinates are independent (students are independent) and assume that they all have the same parameter μ so that X has a density on $\mathbb{R}^5$ given by

$$f(X^{(1)}, \dots, X^{(5)}) = \frac{1}{(2\pi)^{5/2}} \exp\left(\frac{1}{2} \sum_{i=1}^{5} (X^{(i)} - \mu)^2 \right)$$

We are interested in finding the maximum of the function

$$\mu \mapsto \frac{1}{(2\pi)^{5/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{5}(X^{(i)}-\mu)^2\Big)\,.$$

Since $x \mapsto \log(x)$ is increasing, it is equivalent to finding the maximum of

$$g(\mu) = -\frac{5}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{5}(X^{(i)}-\mu)^2\,.$$

We take the derivative and set it zero to get $g'(\hat\mu)=0$ if and only if

$$\hat\mu = \frac{1}{5}\sum_{i=1}^{5} X^{(i)}\,.$$

Therefore, here too, the maximum likelihood estimator of μ is the average $\hat\mu$.

In the case of all three measurements, we are interested in the vector $(\mu_1,\mu_2,\mu_3)\in\mathbb{R}^3$ and it is easy to check that the maximum likelihood estimator is given by $(\hat\mu_1,\hat\mu_2,\hat\mu_3)\in\mathbb{R}^3$ where

$$\hat\mu_j = \frac{1}{5}\sum_{i=1}^{5} X_j^{(i)}$$

4. PERFORMANCE

The next question is: "how accurate is this estimate"? Clearly if we had collected a sample of size 10,000, it would have been more accurate. But what is the effect of the sample size on accuracy?

From here on, we focus on the Gaussian case and always assume that $\sigma^2=1$. Our goal is to assess how good is the performance of the maximum likelihood estimator, that is to measure how large $\hat\mu-\mu$ is. Note that if $X_1,\ldots,X_5$ are random variables then

$$\hat\mu-\mu = \frac{1}{5}\sum_{i=1}^{5} X^{(i)} - \mu$$

is also a random variable and it is desirable to get a global understanding of how large this quantity is.

4.1 Expectation

The expectation (or expected value) of a discrete random variable N is defined as

$$\mathbb{E}[N] = \sum_{k=1}^{n} k\mathbb{P}(N=k)$$

and that of a continuous random variable X with density f is given by

$$\mathbb{E}[X] = \int_{\mathbb{R}} xf(x)dx$$

The operator $\mathbb{E}$ maps a random variable to a real number and enjoys some very nice properties that follow from the property of probabilities. We will use the following ones:

1. For any two random variables X and Y and deterministic real number a: $\mathbb{E}[aX+Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$ (linearity)

2. For any two *independent* random variables X and Y: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Intuitively, the expectation is the average of a random variable and is therefore a measure of location.

If Y is a Bernoulli distribution with parameter p. Then $\mathbb{E}[Y] = 0 \cdot \mathbb{P}(Y=0) + 1 \cdot \mathbb{P}(Y=1) = \mathbb{P}(Y=1) = p$. Moreover, it follows from (2.1) and the linearity of expectation that if $N \sim \mathsf{Bin}(n,p)$ then $\mathbb{E}[N] = np$.

For the Gaussian random variable $X \sim \mathcal{N}(\mu, 1)$ then

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x \exp(-\frac{1}{2}(x-\mu)^2)dx = \mu\,.$$

Moreover, using a change of variables it can be easily checked that $X - \mu \sim \mathcal{N}(0,1)$.

4.2 Quadratic risk

Recall that we are interested in understanding the size of the random variable $\hat{\mu} - \mu$ so a natural candidate is its expectation. However, in this case, it is not hard to see that $\mathbb{E}[\hat{\mu} - \mu] = 0$, regardless of how many students are sampled. The problem is that this measure does not account for the variability of $\hat{\mu}$. Instead, we measure the *quadratic risk* of $\hat{\mu}$ at μ defined by

$$R(\hat{\mu}, \mu) = \mathbb{E}[(\hat{\mu} - \mu)^2]$$

more generally for two random vectors $\hat{\mu}, \mu$, we define

$$R(\hat{\mu}, \mu) = \mathbb{E}[\|\hat{\mu} - \mu\|^2]\,,$$

where $\|\cdot\|$ denotes the Euclidean norm.

In the case of one single measurement across 5 (independent) students, we have

$$R(\hat{\mu}, \mu) = \mathbb{E}[(\hat{\mu} - \mu)^2] = \mathbb{E}\Big[\Big(\frac{1}{5}\sum_{i=1}^{5} Z^{(i)}\Big)^2\Big]$$

where $Z^{(i)} = X^{(i)} - \mu \sim \mathcal{N}(0,1)$ are independent. Using the properties of the expectation, it yields

$$R(\hat{\mu}, \mu) = \frac{1}{25}\sum_{i,j=1}^{5} \mathbb{E}[Z^{(i)}Z^{(j)}] = \frac{1}{25}\sum_{i=1}^{5} \mathbb{E}[(Z^{(i)})^2] = \frac{1}{5}$$

where the last inequality follows from integration by parts: $\mathbb{E}[(Z^{(i)})^2] = 1$.

Note that $R(\hat{\mu}, \mu)$ is independent of μ. More generally, if we have n observations it can be checked that $R(\hat{\mu}, \mu) = 1/n$.

In the case of three body measurements, we have

$$R(\hat{\mu}, \mu) = \mathbb{E}[\|\hat{\mu} - \mu\|^2] = \frac{3}{5}.$$

This seems to be the best we can do since the estimator is so natural. What else could do better?

5. THE STEIN PARADOX

This is perhaps the most surprising result in statistics: there exists an estimator that performs better than the maximum likelihood estimator under the quadratic risk. It is called the *James-Stein estimator*. In the case of three body measurements, it is defined as follows:

$$\tilde{\mu} = \Big(1 - \frac{1}{5\|\hat{\mu}\|^2}\Big)\hat{\mu}\,.$$

where $\hat{\mu}$ is the maximum likelihood estimator. More generally if we had d body measurements measured over n students, the formula would be

$$\tilde{\mu} = \Big(1 - \frac{d-2}{n\|\hat{\mu}\|^2}\Big)\hat{\mu}\,.$$

In effect, this estimator is *shrinking* the maximum likelihood estimator.

We are now going to show that $R(\tilde{\mu}, \mu) < R(\hat{\mu}, \mu)$ for all $\mu \in \mathbb{R}^3$ as long as the body measurements are independent of each other (which may be questionable).

To that end, observe that

$$\begin{aligned} R(\tilde{\mu}, \mu) &= \mathbb{E}\Big[\Big\|\Big(1 - \frac{1}{5\|\hat{\mu}\|^2}\Big)\hat{\mu} - \mu\Big\|^2\Big] \\ &= \mathbb{E}\Big[\Big\|\hat{\mu} - \mu - \frac{\hat{\mu}}{5\|\hat{\mu}\|^2}\Big\|^2\Big] \\ &= \mathbb{E}\Big[\|\hat{\mu} - \mu\|^2\Big] + \mathbb{E}\Big[\frac{1}{25\|\hat{\mu}\|^2}\Big] - \frac{2}{5}\sum_{j=1}^{3}\mathbb{E}\Big[\frac{(\hat{\mu}_j - \mu_j)\hat{\mu}_j}{\|\hat{\mu}\|^2}\Big]\,. \end{aligned}$$

Since $\hat{\mu}_j \sim \mathcal{N}(\mu_j, 1/5)$, we have

$$\mathbb{E}\Big[\frac{(\hat{\mu}_j - \mu_j)\hat{\mu}_j}{\|\hat{\mu}\|^2}\Big] = \Big(\frac{5}{2\pi}\Big)^{3/2}\iiint_{\mathbb{R}^3}\frac{(x_j - \mu_j)x_j}{x_j^2 + x_2^2 + x_3^2}\exp\Big[-\frac{5}{2}\sum_{i=1}^{3}(x_i - \mu_i)^2\Big]dx_1dx_2dx_3$$

To apply an integration by parts notice that if we look only at the integral with respect to x_j, we have

$$\frac{\partial}{\partial x_j}\frac{x_j}{x_j^2 + x_2^2 + x_3^2} = \frac{\|x\|^2 - 2x_j^2}{\|x\|^4}$$

and

$$\frac{\partial}{\partial x_j}\Big\{-\frac{1}{5}\exp\Big[-\frac{5}{2}\sum_{i=1}^{3}(x_i - \mu_i)^2\Big]\Big\} = (x_j - \mu_j)\exp\Big[-\frac{5}{2}\sum_{i=1}^{3}(x_i - \mu_i)^2\Big]$$

Using integration by parts we get

$$\begin{aligned} \mathbb{E}\Big[\frac{(\hat{\mu}_j - \mu_j)\hat{\mu}_j}{\|\hat{\mu}\|^2}\Big] &= \frac{1}{5}\Big(\frac{5}{2\pi}\Big)^{3/2}\iiint_{\mathbb{R}^3}\frac{\|x\|^2 - 2x_j^2}{\|x\|^4}\exp\Big[-\frac{5}{2}\sum_{i=1}^{3}(x_i - \mu_i)^2\Big]dx_1dx_2dx_3 \\ &= \frac{1}{5}\mathbb{E}\Big[\frac{\|\hat{\mu}\|^2 - 2\hat{\mu}_j^2}{\|\hat{\mu}\|^4}\Big] \\ &= \frac{1}{5}\mathbb{E}\Big[\frac{1}{\|\hat{\mu}\|^2}\Big] - \frac{2}{5}\mathbb{E}\Big[\frac{\hat{\mu}_j^2}{\|\hat{\mu}\|^4}\Big] \end{aligned}$$

Summing up over j yields

$$\sum_{j=1}^{3} \mathbb{E}\Big[\frac{(\hat{\mu}_j - \mu_j)\hat{\mu}_j}{\|\hat{\mu}\|^2}\Big] = \Big(\frac{3}{5} - \frac{2}{5}\Big)\mathbb{E}\Big[\frac{1}{\|\hat{\mu}\|^2}\Big] = \frac{1}{5}\mathbb{E}\Big[\frac{1}{\|\hat{\mu}\|^2}\Big]$$

Therefore

$$R(\tilde{\mu}, \mu) = R(\hat{\mu}, \mu) - \frac{1}{25}\mathbb{E}\Big[\frac{1}{\|\hat{\mu}\|^2}\Big] < R(\hat{\mu}, \mu)$$

6. EXERCISES

6.1 Problem 1

Prove that in the case $n = 1$ but d is general, the James-Stein estimator outperforms the maximum likelihood estimator in the Gaussian case if and only if $d \geq 3$.

6.2 Problem 2

Fix $\lambda > 0$ and let X be a random variable with density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

1. Compute $\mathbb{E}[X]$
2. Compute $g(t) = P(X \leq t)$ for any $t > 0$.
3. Show that $g'(t) = f(t)$
4. Let $X_1, \ldots, X_n$ be n independent random variables with the same distribution as X. Find the density of $Y = \max_i X_i$.
5. Compute the maximum likelihood estimator of λ from the observations $X_1, \ldots, X_n$.