Mean and sd of null distribution

Sayani Gupta

Question:1

I'm definite that the theeretical distribution of Jensen Shannon statistic follows a Chi-squared distribution with m degrees of freedom. If m is sufficiently large, Chi square distribution will tend to a Normal distribution with mean m and variance $2m$.

Empirical test: In my case, do we mean that no matter what is the underlying distribution of the dataset, if you compute Jensen Shannon distances they will be normal like?

Let's find out!

Simulation study:

1. Fix n.
2. Take a particular distribution and generate N random samples of size n each.
3. Compute pairwise distances for each random sample.
4. Construct a histogram to see of the pairwise distances follow Normal?

Question:2

Let's suppose they are normal! How would the max of these normal distances behave? Would the distribution of the maximum differ if you are taking max between 2 random variables or maximum between 10 random variables?

Yes they would! So what do you do about it?

We want to make the distribution of the max same for any n. By making the distribution same, we just mean that at least the mean and sd of the distribution of the maximum should be same for varied n. This means we should be able to compare values across different n. What is a consistent estimator of max?

Simulation Study

Aim: To see if norm_max is working after a certain n for all cases. 1. Generate N(1:10,10) distribution and compute distribution of max and norm_max.
2. Generate N(10,1:10) distribution and compute distribution of max and norm_max.

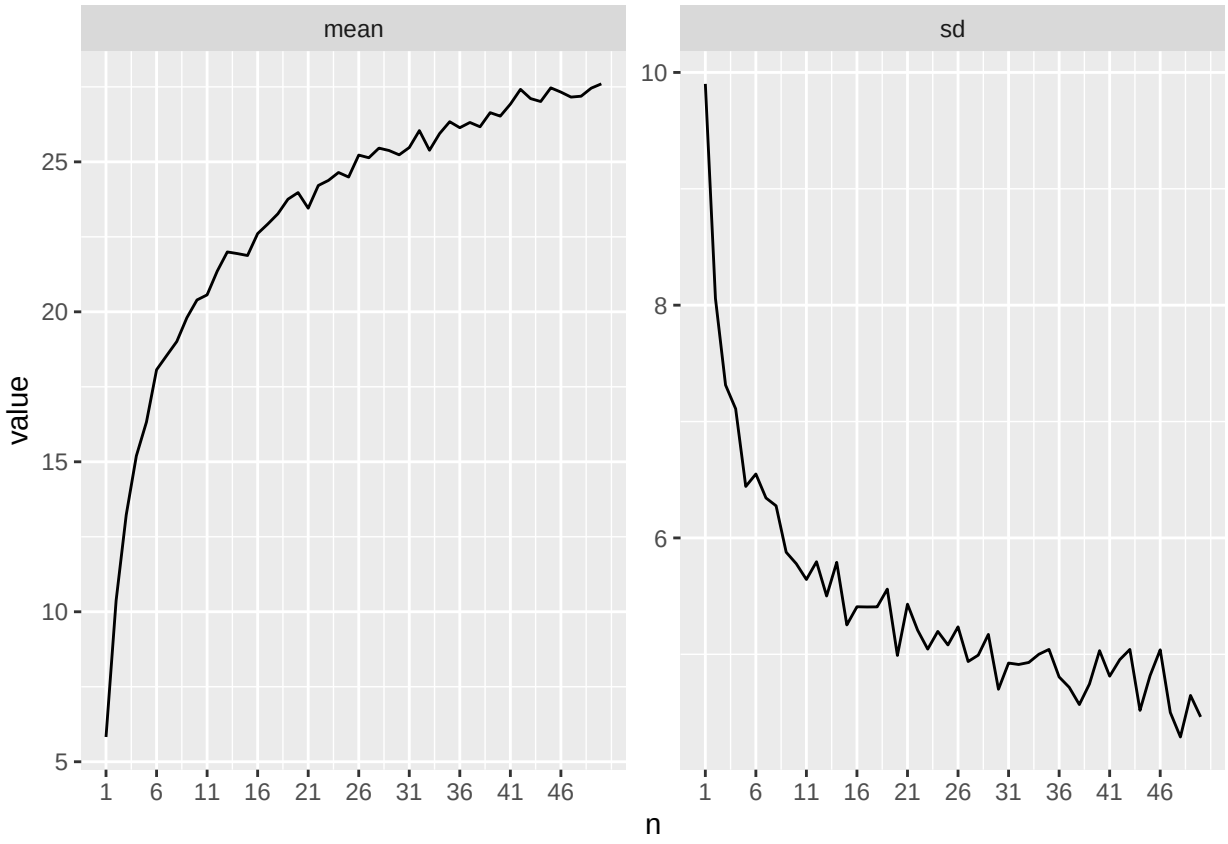
Behavior of mean and sd of the distribution of maximum

Suppose that X_1, X_2 , ... , X_n are i.i.d. random variables with expected values $E(X_i) = \mu < \infty$ and variance $Var(X_i) = \sigma^2 < \infty$. Let $Y = max(X_1, X_2, \ldots, X_n)$.

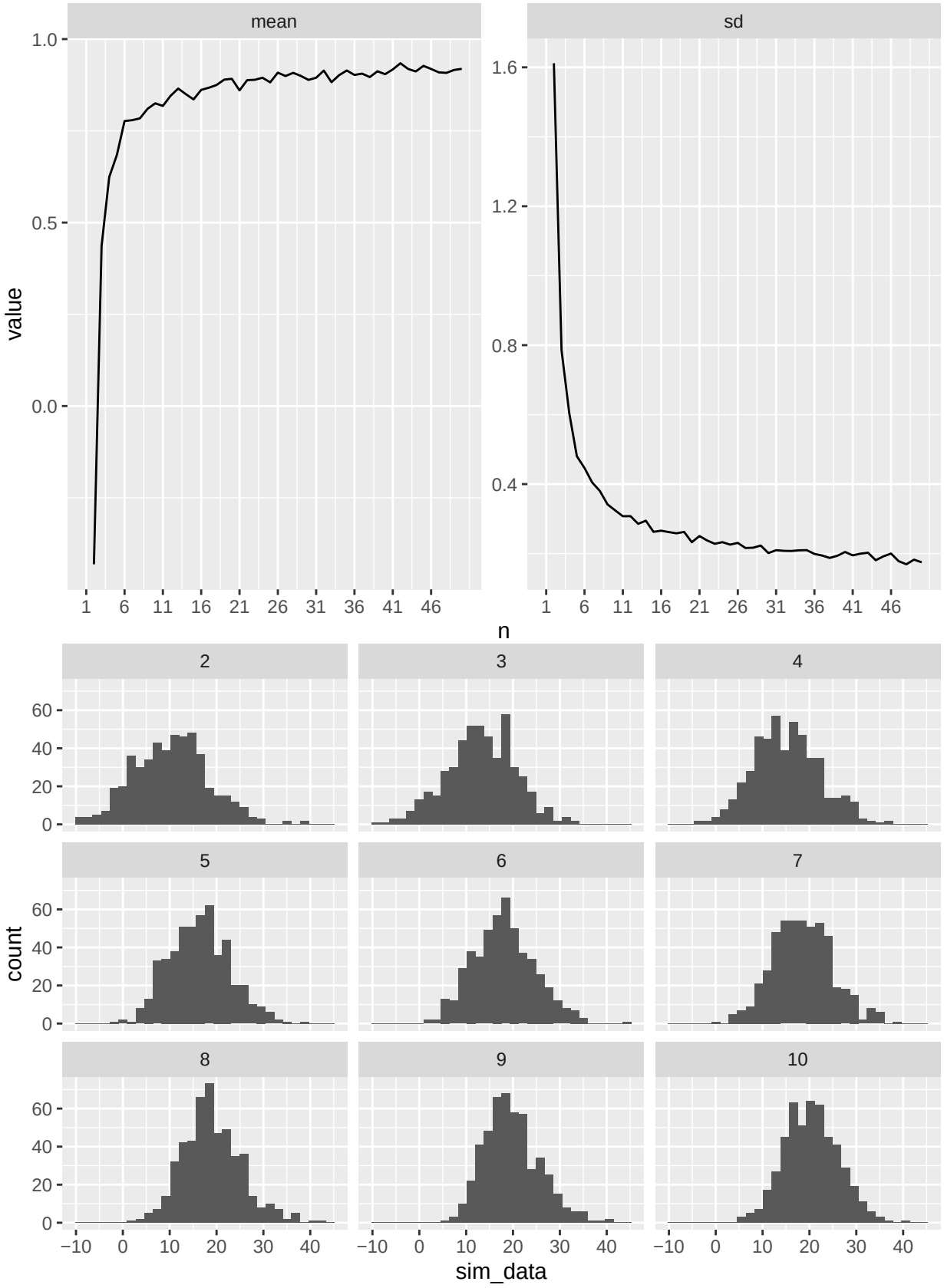
Let $F_X(x)$ be the common distribution of the variables X_i and let $F_Y(y)$ be the corresponding distribution of Y. $F_Y(y)$ could be obtained from $F_X(x)$ simply by using: $F_Y(y) = P[(X_1 \leq y) \cap (X_2 \leq y) \cap ... \cap (X_n \leq$

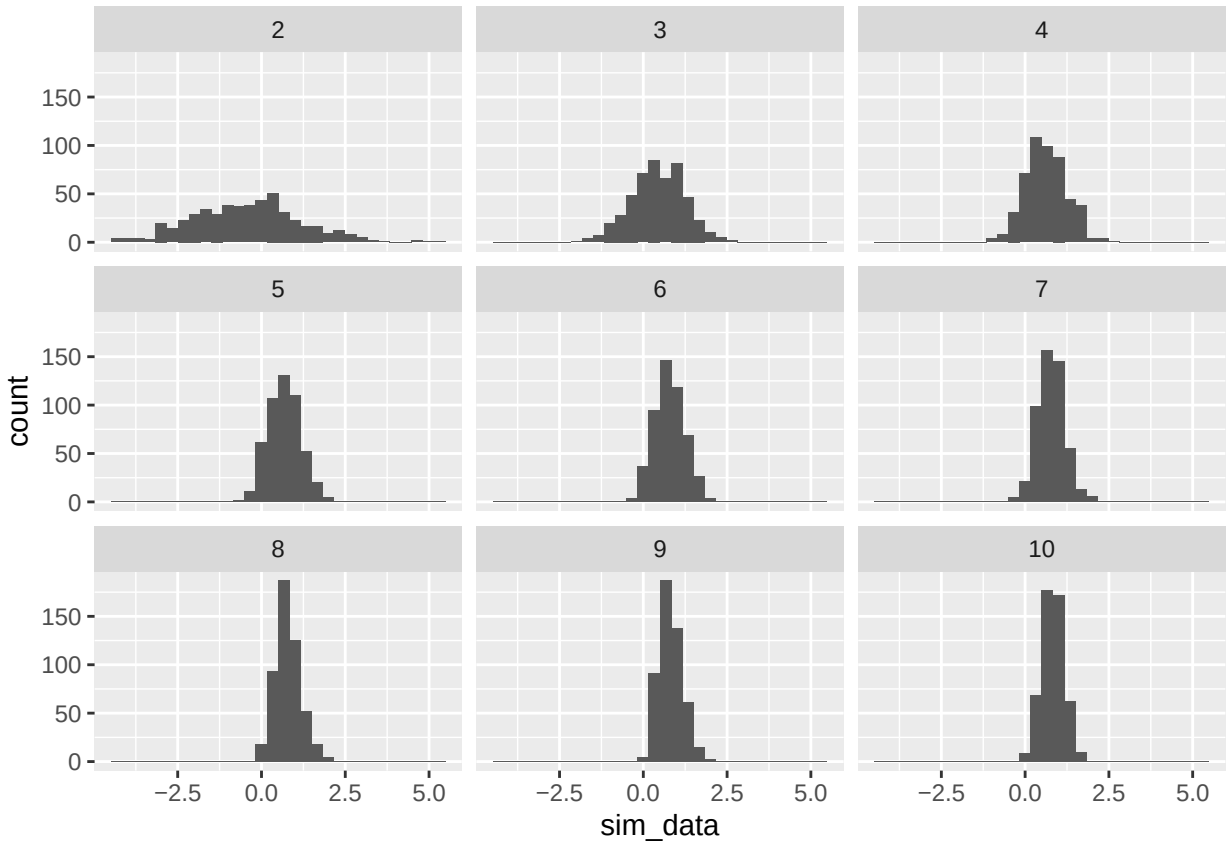
$y)] = F_X(y)^n$. For large n, the distribution of Y approaches a standard shape, which does not depend on F_X. But what about the case when n is not large enough? The distribution of maximum in that case will indeed depend on n and the underlying distribution of X. If $F_X(x)$ is the CDF of X, then $F_Y(y) = F_x(y)^n$. Suppose Φ nd ϕ are the cdf and pdf of a standard normal distribution, then $f_Y(y) = n\Phi(y)^{n-1}\phi(y)$, which depends on n. Hence, we are trying to normalise for n. Also, it depends on the underlying distribution of X, which we have assumed as normal in our case. As n grows, we can see the right tail growing, which implies that the probability that we will get a higher maximum is more. Now, for large n, we used EVT to normalise for n, that is, we brought them to the same scale without distorting the range of the distribution. But in our case, we will mostly have small n. It is important to ensure that they have the same mean and variation, for being able to compare the maximum value across n. We observe from the following graphs that our normalisation works after $n = 6$, after which the difference in mean and standard deviation flattens out a lot.

Mean and standard deviation of the distribution of maximum



Mean and standard deviation of the distribution of normalised maximum





Looking at the smaller values of n, what we already saw last week.

Distribution of max for smaller n

Distribution of normalised max for smaller n

Aim: To compute mean and sd by bootstrap method for small samples and compute norm_max to see if it is working for n cases.

Non-parametric bootstrap

Assuming a data set $x = (x_1, ..., x_n)$ is available.
1. Fix the number of bootstrap re-samples N. Often $N \in [1000, 2000]$.
2. Sample a new data set x^* set of size n from x with replacement (this is equivalent to sampling from the empirical cdf $\hat{F}$).
3. Estimate θ from x^*. Call the estimate $\hat{\theta}$ $\forall i \in 1, \ldots, N$. Store.
4. Repeat step 2 and 3 N times.
5. Consider the empirical distribution of $(\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_N^*)$ as an approximation of the true distribution of $\hat{\theta}$.

We assume that distances are normal, but we do not assume anything about the distribution of max distances. So using the non-paramteric bootstrap approach to estimate the $\theta = (\mu, \sigma)$.

I construct the data structure for $n = 2$ here to ensure that the methodology is correct.

N	x_1	x_2	max
1	$x_{1,1}$	$x_{1,2}$	m_1
2	.	.	m_2
.	.	.	.
.	.	.	.
N	$x_{1,N}$	$x_{2,N}$	m_N

Question:3

This is just for one level! What do you do for the other level?

Do you take median or max. Given that you take median how do you normalise for the value of n there? What is a consistent estimator of median?