Randomness in Correlation and Correlation Hacking: Explicit Distributions

Summary : Let $X_1, X_2, \ldots X_n$ and $Y_1, Y_2, \ldots Y_n$ be independent Gaussian Standard *r.v.s.* with $\mathbb{E}\left(\sum_{i=1}^{n} X_i Y_i\right) = 0$ .We are looking for the exact distributions $\psi(\rho_n)$ and $\varphi\left(\rho_{n,d}^{\max}\right)$ where $\rho_n^{\max} = \sup(\rho_{n,i})$, $i = 1, 2, \ldots d$, d being the number of separate realizations of the pair (X, Y).

$$\rho_n := \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\ \sqrt{\sum_{i=1}^{n} Y_i^2}},\ n > 1 \in \mathbb{N},\ \mathbb{E}(\rho_n) = 0 \text{ for all } n$$

Theorem :

For n>0 even,

$$\psi(\rho, n) = \begin{cases} -\dfrac{\sqrt{\pi}\ \sec\left(\frac{\pi n}{2}\right)\left(\rho\,\Gamma\left(\frac{n+1}{2}\right)^2\ {}_2\tilde{F}_1\left(\frac{n+1}{2}, \frac{n+1}{2}; \frac{3-n}{2}; \rho^2\right) - \Gamma(n)^2\ \mathrm{sgn}(\rho)\ |\rho|^n\ {}_2\tilde{F}_1\left(n, n; \frac{n+1}{2}; \rho^2\right)\right)}{\rho\,\Gamma\left(\frac{n}{2}\right)^3} & \rho \in [-1, 1)\setminus 0 \\ 0 & \text{elsewhere} \end{cases}$$

Background: How significant, i.e. nonrandom, is a correlation coefficient (Pearson's product moment) for a bivariate normal distribution? For random matrices, we have a well-developed Marchenko-Pastur theory. Here we are confronted with a simple problem: pairwise correlation. **Understanding the derivations will allow us to generalize and obtain the distribution of the maximum (or minimum) used in p-hacking.**

The mission for us to ***always rederive***, as many things taken for granted are often misapplied. This exercise for instance will show that the usual approximation $\phi(\rho_n)$ (for the case E $(\rho) = 0$ for all n) inspired from an old derivation by Fisher in 1912, does not work well for small *n.*

$$\phi(\rho_n) := \frac{\left(1 - \rho_n^2\right)^{\frac{n-4}{2}}}{\mathrm{B}\left(\frac{1}{2}, \frac{n-2}{2}\right)},\ \text{where } B(.,.) \text{ is the Beta Function}$$

There are many routes (one entailing cosine of angles). We opt for the Fourier route.

$$\text{Let } \rho_n := \frac{Z_n}{W_n},\ Z_n = \sum_{i=1}^{n} X_i\, Y_i\,,\ \ W_n = \sqrt{\sum_{i=1}^{n} X_i^2}\ \sqrt{\sum_{i=1}^{n} Y_i^2}$$

Step 0: Verifying that Z_{100} and W_{100} are independent by Monte Carlo Confirmation

```
tameta = Table[ta = Table[(a = RandomVariate[NormalDistribution[], 100];
      b = RandomVariate[NormalDistribution[], 100]);
    {a.b, Sqrt[Total[a^2]] Sqrt[Total[b^2]]}, {10^3}] // Transpose;
  {KendallTau[ta[[1]], ta[[2]]], SpearmanRho[ta[[1]], ta[[2]]],
   Correlation[ta[[1]], ta[[2]]], HoeffdingD[ta[[1]], ta[[2]]]}, {10^3}];
```

```
Mean[tameta]
```

```
{-0.0000552032, -0.000152817, -0.000272621, 0.000806137}
```

Step1: distribution of Z_n

The distribution $Z_n = \sum_{i=1}^{n} X_i\, Y_i$ has for PDF

$$f_1 = \frac{2^{\frac{1}{2}-\frac{n}{2}}\,\sigma^{-1-n}\,\text{Abs}[x]^{\frac{1}{2}(-1+n)}\,\text{BesselK}\left[\frac{1-n}{2}, \frac{\text{Abs}[x]}{\sigma^2}\right]}{\sqrt{\pi}\ \text{Gamma}\left[\frac{n}{2}\right]},\quad x \in (-\infty, \infty)$$

Proof. The Characteristic function $\chi_x(t) = e^{-\frac{1}{2} t^2 \sigma^2}$, $f_x(x)$= PDF = $\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}$

There is a theorem showing that the chf of a product X and Y(both independent) is

$$\chi_{x\,y}(t) = \int_{-\infty}^{\infty} \chi_x(t\,y)\, f_y(y)\, dy = \int_{-\infty}^{\infty} \chi_y(t\,x)\, f_x(x)\, dy$$

Since both X and Y follow the same distribution:

Integrate: $\chi_z(t) \int_{-\infty}^{\infty} \chi_x(t\,x)\, f(x)\, dx = \frac{1}{\sqrt{\sigma^4\, t^2 + 1}}$

The Characteristic function of n-summed variables: $\chi_{z_n}(t) = \left(1 + t^2\, \sigma^4\right)^{-n/2}$

Hence by the inversion theorem:

$$f_{Z_n(Z)} = \frac{1}{2\,\pi} \int_{-\infty}^{\infty} \exp(-i\, t\, z) \left(\frac{1}{\sqrt{\sigma^4\, t^2 + 1}}\right)^n dt = \frac{2^{\frac{1}{2}-\frac{n}{2}}\,\sigma^{-1-n}\,\text{Abs}[z]^{\frac{1}{2}(-1+n)}\,\text{BesselK}\left[\frac{1-n}{2}, \frac{\text{Abs}[z]}{\sigma^2}\right]}{\sqrt{\pi}\ \text{Gamma}\left[\frac{n}{2}\right]}$$

Step 2a: distribution W_n^2

The distribution $W_n^2 = \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2$ has for PDF

$$f_{2_a} = \frac{2^{1-n}\, z^{-1+\frac{n}{2}}\,\text{BesselK}\left[0, \sqrt{z}\right]}{\Gamma\left[\frac{n}{2}\right]^2},\quad z > 0$$

Instead of going directly to W_n we operate in 2 steps.

Proof: By a standard result,

$\sum_{i=1}^{n} X_i^2$ **≈ Chi Square(n) ,** same for $\sum_{i=1}^{n} Y_i^2$. Both sums are independent.

We need the product.

Characteristic function of $\sum_{i=1}^{n} X_i^2$

$$\Xi^2(n) = (1 - 2\,i\,t)^{-n/2}$$

Characteristic function of $w = \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2$

$$\chi_W(t) = \int_0^\infty \frac{2^{-\frac{n}{2}} e^{-\frac{w}{2}} w^{\frac{n}{2}-1} (1 - 2\,i\,t\,w)^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\, dw = 2^{-n} (-i\,t)^{-n/2}\, \text{HypergeometricU}\left[\frac{n}{2}, 1, \frac{i}{4\,t}\right]$$

Step 2b: W_n

The distribution $W_n = \sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}$ has for PDF

$$f_2 = \frac{2^{2-n}\, z_2^{-1+n}\, \text{BesselK}[0, z_2]}{\Gamma\left[\frac{n}{2}\right]^2}, \quad z_2 > 0$$

Proof: straight transformation of r.v.

Step 3: Distribution of ratio Z_n / W_n

Given independence, the joint distribution is

$$f_1 f_2 = \frac{2^{\frac{5}{2}-\frac{3n}{2}}\, z_2^{-1+n}\, \text{Abs}[x]^{\frac{1}{2}(-1+n)}\, \text{BesselK}[0, z_2]\, \text{BesselK}\left[\frac{1}{2}(-1+n), \text{Abs}[x]\right]}{\sqrt{\pi}\, \Gamma\left[\frac{n}{2}\right]^3}$$

This is more involved. Let $r = \frac{x}{z_2}$, $w = z_2$ we have by standard change of random variable

$$f(\rho, w) = \frac{2^{\frac{5}{2}-\frac{3n}{2}}\, \text{Abs}[\rho\, w]^{-\frac{1}{2}+\frac{n}{2}}\, \text{BesselK}[0, w]\, \text{BesselK}\left[\frac{1}{2}(-1+n), \text{Abs}[\rho]\, w\right] w^n}{\sqrt{\pi}\, \Gamma\left[\frac{n}{2}\right]^3}$$

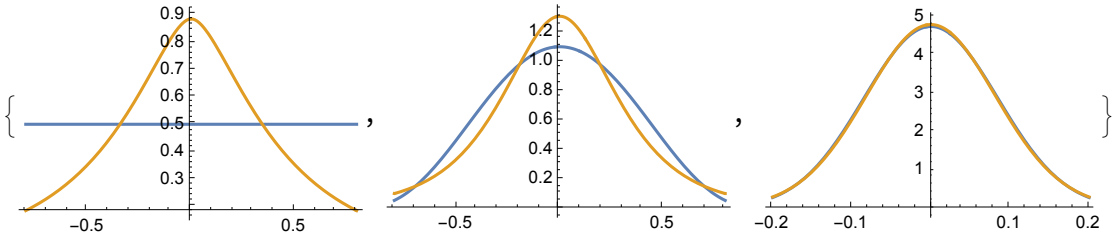
Finally, the marginal $\psi(\rho) = \int f(\rho, w)\, dw$. (Shortcut: we write ρ for ρ_n)

For n even:

$$\psi(\rho) = \begin{cases} -\dfrac{\sqrt{\pi}\, \sec\left(\frac{\pi n}{2}\right)\left(\rho\, \Gamma\left(\frac{n+1}{2}\right)^2\, {}_2\tilde{F}_1\left(\frac{n+1}{2}, \frac{n+1}{2}; \frac{3-n}{2}; \rho^2\right) - \Gamma(n)^2\, \text{sgn}(\rho)\, |\rho|^n\, {}_2\tilde{F}_1\left(n, n; \frac{n+1}{2}; \rho^2\right)\right)}{\rho\, \Gamma\left(\frac{n}{2}\right)^3} & \rho \in [-1, 1) \setminus 0 \\ 0 & \text{elsewhere} \end{cases}$$

For n odd we can do some more complicated work.

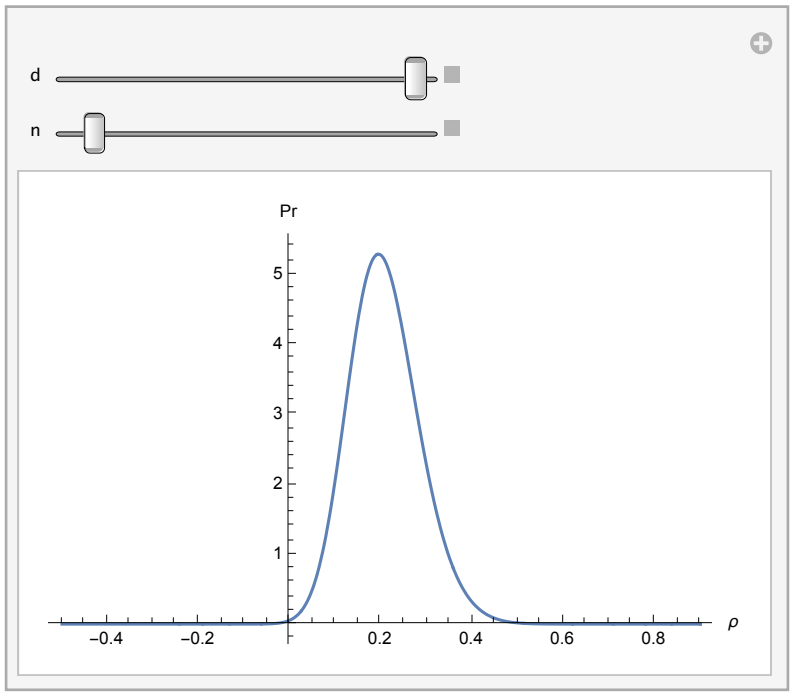
We can see for 4, 10, and 144 how the blue approximation fails to properly capture the distribution



In[1]:=
```
Clear[distmax];
distmax[ρ_, n_, d_] := d (1 - ρ^2)^(-2 + n/2) ((Sqrt[π] Gamma[-1 + n/2])/(2 Gamma[1/2 (-1 + n)]) + ρ Hypergeometric2F1[1/2, 2 - n/2, 3/2, ρ^2])^(-1 + d) / Beta[1/2, 1/2 (-2 + n)];
```

Distribution of maximum spurious correlation with d variables and *n* data points per variable. Shows the expected bias from “data hacking” when the operator has d pairs to choose from.

```
Manipulate[Plot[distmax[ρ, n, d], {ρ, -.5, .9},
  PlotRange → All, AxesLabel → {ρ, "Pr"}], {d, 4, 10, 1}, {n, 4, 1000, 2}]
```



- Expected Max

Discrete Variables.