HLMA408: Traitement des données

Loi normale / gaussienne

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

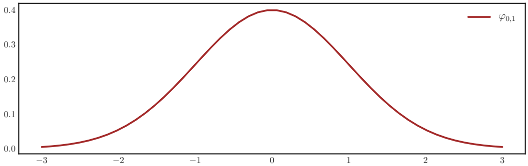
Sommaire

Loi normale standard (ou centrée-réduite)

- Une v.a. réelle X suit une "**loi normale**" ou "**loi gaussienne**" ou "loi de Laplace-Gauss" si sa densité ((: *probability density function (pdf)*)) vaut, $\forall x \in \mathbb{R}$:

$$\boxed{\varphi(x) = \varphi_{0,1}(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}$$

- Notation: $X \sim \mathcal{N}(0, 1)$



- Propriétés: $\mathbb{E}(X) = 0$ (espérance nulle) et $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = 1$ (variance unitaire)

Loi normale

- Une v.a. Y suit une loi normale de paramètres μ et σ^2 si

$$Y = \mu + \sqrt{\sigma^2} X$$

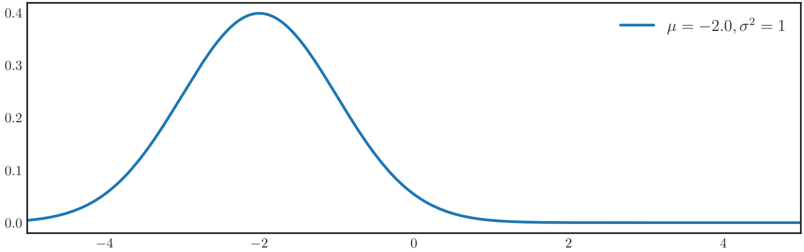
  où $X \sim \mathcal{N}(0,1)$, c'est-à-dire si sa densité vaut:

$$\boxed{\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

- <u>Notation</u>: $Y \sim \mathcal{N}(\mu, \sigma^2)$

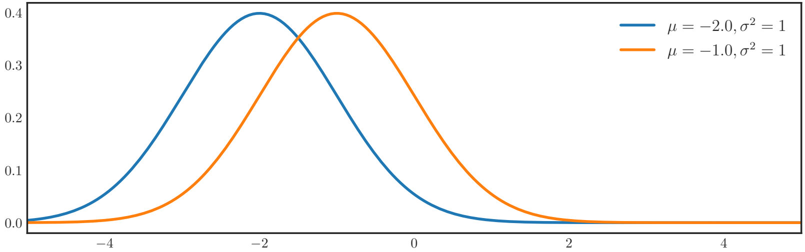
- <u>Propriétés</u>: $\mathbb{E}(Y) = \mu$ et $\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y - \mathbb{E}(Y))^2 = \sigma^2$

Exemple: variation selon μ (centrage)



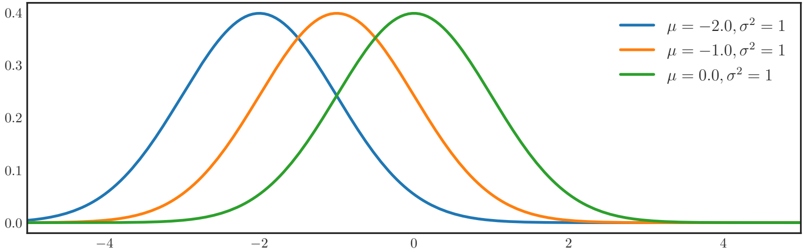
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon μ (centrage)



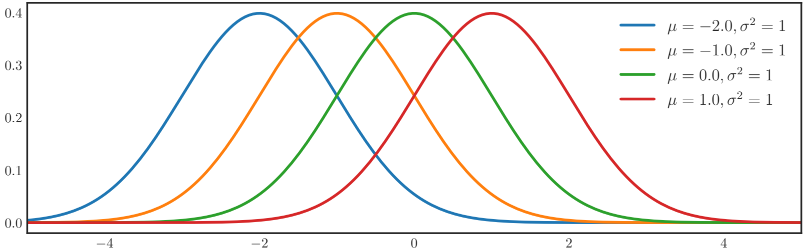
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon μ (centrage)



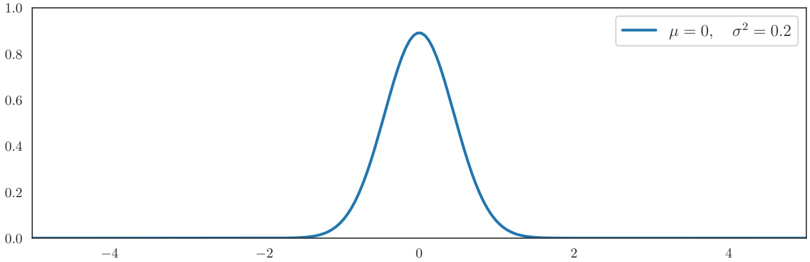
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon μ (centrage)



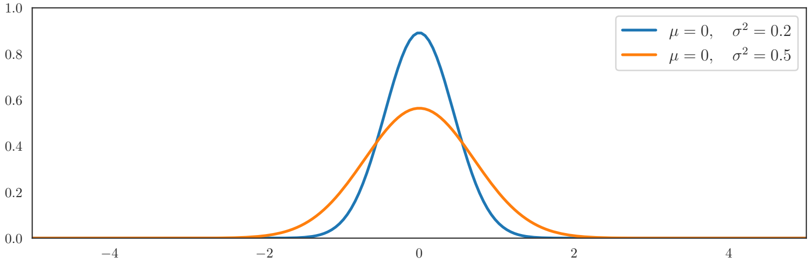
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon σ (échelle)



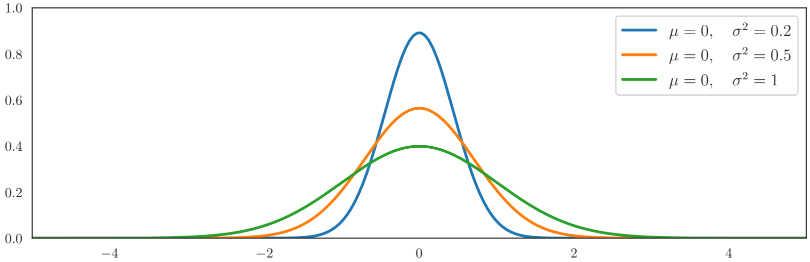
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon σ (échelle)



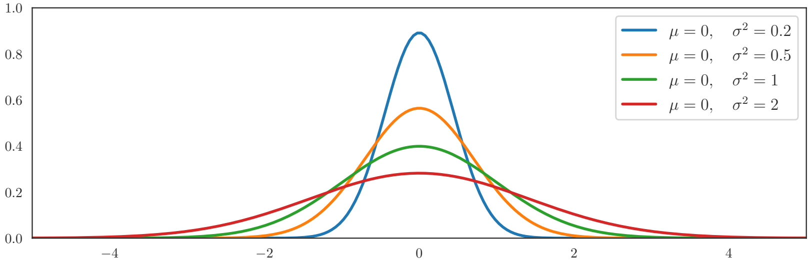
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon σ (échelle)



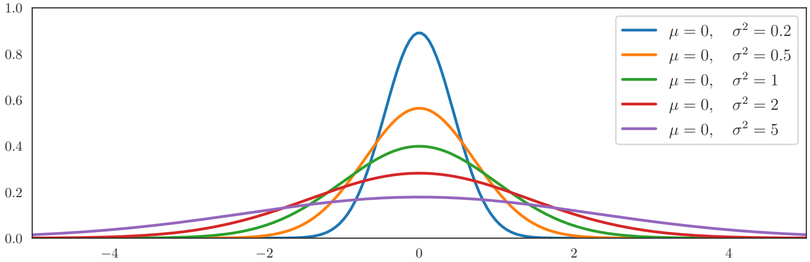
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Exemple: variation selon σ (échelle)



TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

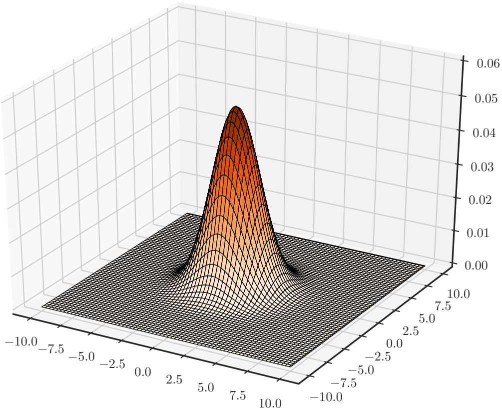
Exemple: variation selon σ (échelle)



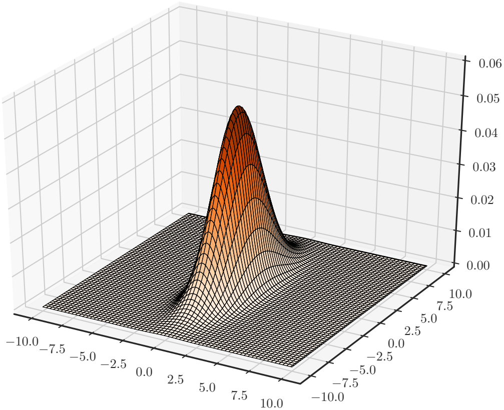
TO DO: voir jupyter notebook `GaussianDistribution.ipynb`

Sommaire

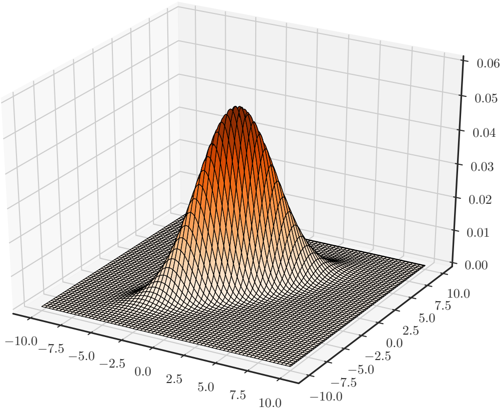
Exemple 2D ($p = 2$)



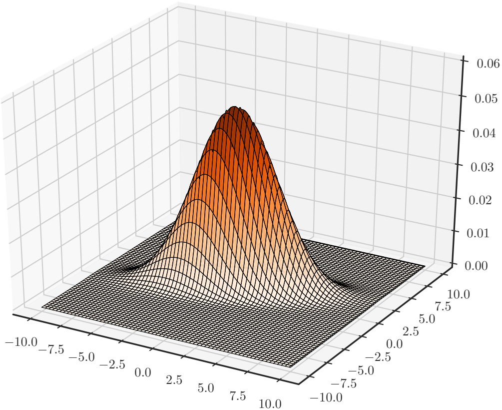
Exemple 2D ($p = 2$)



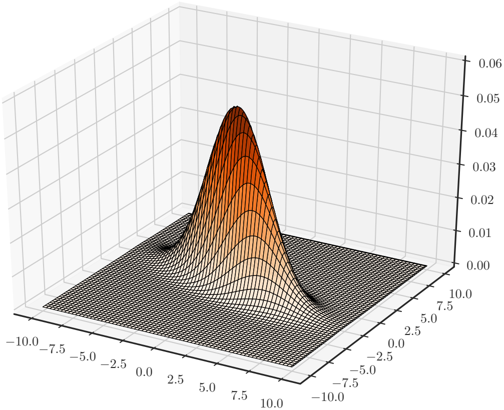
Exemple 2D ($p = 2$)



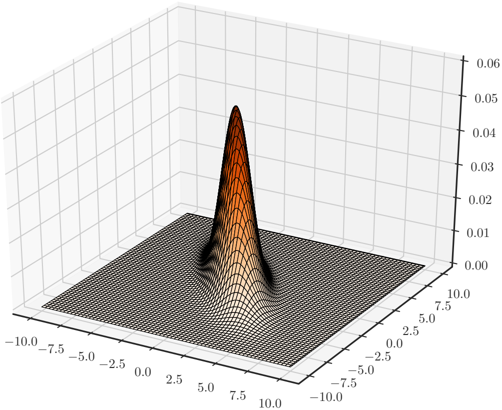
Exemple 2D ($p = 2$)



Exemple 2D ($p = 2$)



Exemple 2D ($p = 2$)



Vecteurs gaussiens (hors programme)

En dimension p, les lois gaussiennes ont des densités de la forme:

$$\boxed{\varphi_{\mu,\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}}\,.$$

$\varphi_{\mu,\Sigma}$: gouvernée par deux paramètres

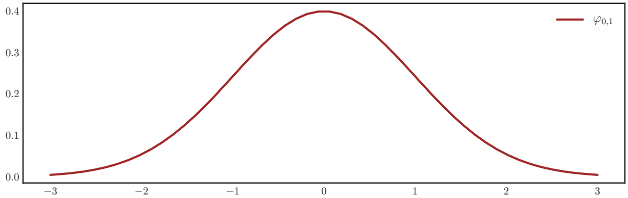
- le vecteur d'espérance $\mu \in \mathbb{R}^p$
- la matrice de covariance $\Sigma \in \mathbb{R}^{p\times p}$ est symétrique

<u>Notation</u>: si le vecteur aléatoire X suit une loi normale d'espérance μ et de covariance Σ (supposée définie positive, *i.e.*, toutes ses valeurs propres sont strictement positives), on note $X \sim \mathcal{N}(\mu, \Sigma)$

<u>Rem</u>: $|\Sigma| = \det(\Sigma)$ est le produit des valeurs propres de Σ. On parle de cas dégénéré quand $\det(\Sigma) = 0$

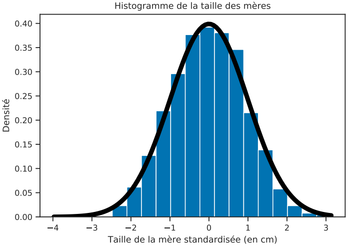
Loi normale

- Rôle central en statistique
- De nombreuses données suivent (approx.) cette loi
- Le théorème central limite (TCL) assure que certaines variables aléatoires suivent (approx.) cette loi si n est grand



Lien histogramme–densité et TCL

- Si des données suivent approximativement une loi normale, alors l'histogramme des données centrées réduites doit ressembler à la courbe ci-dessous
- Standardiser les données $x_1, \ldots, x_n$: $\dfrac{x_i - \bar{x}_n}{s_n}, \quad i = 1, \ldots, n$ (retirer la moyenne, diviser par l'écart-type)



Comparaison: histogramme / loi normale

- Les données semblent être bien représentées par une loi normale
- On peut alors utiliser cette loi pour répondre à des questions statistiques

Calcul des probabilités

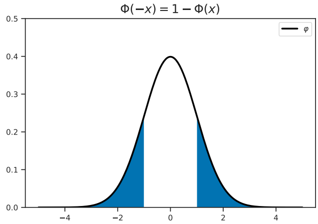
- La probabilité d'être plus petit qu'un nombre z correspond à l'aire sous la courbe φ entre $-\infty$ et z

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Φ est la **fonction de répartition** (: *Cumulative distribution function (cdf)*) d'une loi normale

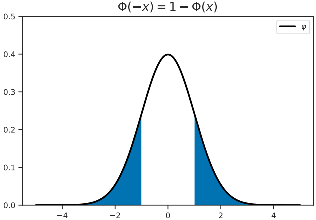
Quelques propriétés de Φ

- $\Phi(-x) = 1 - \Phi(x)$ (symétrie):

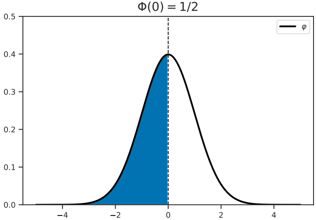


Quelques propriétés de Φ

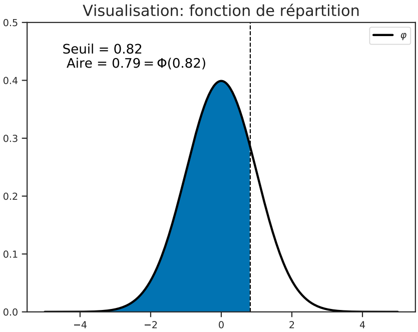
- $\Phi(-x) = 1 - \Phi(x)$ (symétrie):



- $\Phi(0) = \frac{1}{2}$ (0 est la médiane):



Visualisation de la fonction de répartition



Exemple de la taille de la mère:

$$\mathbb{P}[\text{Taille} \leqslant 168] = \mathbb{P}\left[\frac{\text{Taille} - \bar{x}_n}{s_n} \leqslant \frac{168 - \bar{x}_n}{s_n}\right] \approx \Phi(0.82) = 0.79$$

on calcule la moyenne ($\bar{x}_n = 162.7$) et l'écart-type ($s_n = 6.428$) de l'échantillon pour obtenir ce nombre

Rem: *cf.* notebook `GaussianDistribution.ipynb`

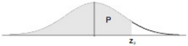


TABLE C.1. Cumulative normal distribution—values of P corresponding to z_p for the standard normal curve.

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.719	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8707	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

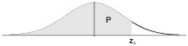


TABLE C.1. Cumulative normal distribution—values of P corresponding to z_p for the standard normal curve.

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	[illegible]	[illegible]	.6293	.6331	.6368	.6406	.6443	[illegible]	.6517
.4	.6554	[illegible]	[illegible]	.6664	.6700	.6736	.6772	[illegible]	[illegible]	.6879
.5	.6915	[illegible]	[illegible]	.7019	.7054	.7088	.7123	[illegible]	.719	.7224
.6	.7257	[illegible]	[illegible]	.7357	.7389	.7422	[illegible]	.7486	.7517	.7549
.7	.7580	[illegible]	[illegible]	.7673	.7704	[illegible]	[illegible]	.7794	.7823	.7852
.8	.7881	[illegible]	[illegible]	.7967	.7995	[illegible]	[illegible]	.8078	.8106	.8133
.9	.8159	[illegible]	[illegible]	[illegible]	.8264	[illegible]	.8315	.8340	.8365	.8389
1.0	.8413	.8438	[illegible]	[illegible]	[illegible]	[illegible]	.8554	.8577	.8599	.8621
1.1	.8643	.8665	[illegible]	[illegible]	[illegible]	.8749	.8707	.8790	.8810	.8830
1.2	.8849	.8869	[illegible]	[illegible]	[illegible]	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	[illegible]	[illegible]	[illegible]	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	[illegible]	[illegible]	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	[illegible]	[illegible]	[illegible]	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	[illegible]	[illegible]	[illegible]	[illegible]	.9515	.9525	.9535	.9545
1.7	.9554	.9564	[illegible]	.9582	[illegible]	[illegible]	.9608	.9616	.9625	.9633
1.8	.9641	[illegible]	[illegible]	.9664	[illegible]	[illegible]	[illegible]	.9693	.9699	.9706
1.9	.9713	[illegible]	[illegible]	.9732	[illegible]	[illegible]	[illegible]	.9756	.9761	.9767
2.0	.9772	[illegible]	[illegible]	.9788	.9793	[illegible]	[illegible]	.9808	.9812	.9817
2.1	.9821	[illegible]	[illegible]	.9834	.9838	[illegible]	[illegible]	.9850	.9854	.9857
2.2	[illegible]	[illegible]	.9868	.9871	.9875	.9878	[illegible]	.9884	.9887	.9890
2.3	[illegible]	[illegible]	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	[illegible]	[illegible]	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	[illegible]	[illegible]	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Quantiles gaussiens

Utiliser plutôt:

en Python

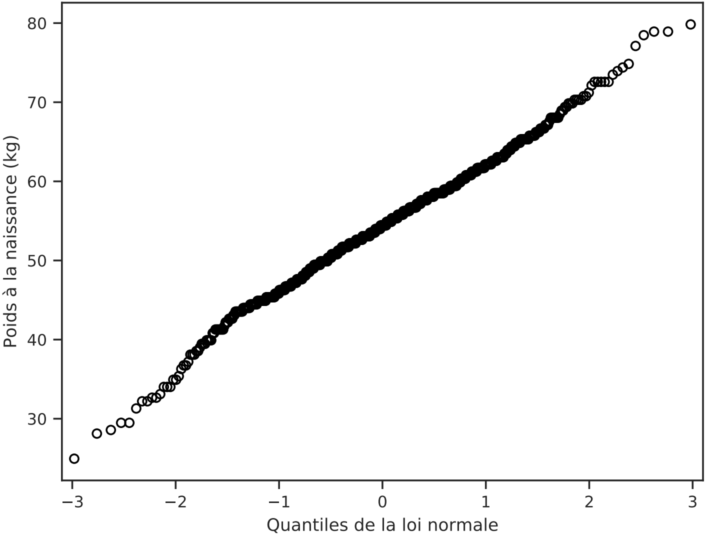
```
>>> from scipy.stats import norm # import
>>> norm.ppf(0.95, 0, 1)
1.6448536269514722
```

en R:

```
>>> qnorm(.95,mean=0,sd=1)
1.6448536269514722
```

Sommaire

Diagramme quantile-quantile[1]: exemple



[1] M. B. Wilk and R. Gnanadesikan. “Probability plotting methods for the analysis for the analysis of data”. In: *Biometrika* 55.1 (1968), pp. 1–17.

Diagrammes quantiles-quantiles (qq-plots)

- Représentation graphique comparant des distributions de type:
  - observées vs observées
  - observées vs théoriques
  - théoriques vs théoriques
- Utilité des qq-plots:
  - Vérifier si les données suivent une loi particulière
  - Vérifier si deux jeux de données ont la même loi
- Construction pour le cas gaussien: on ordonne l'échantillon $x_1, \ldots, x_n$ en $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ et on affiche les points

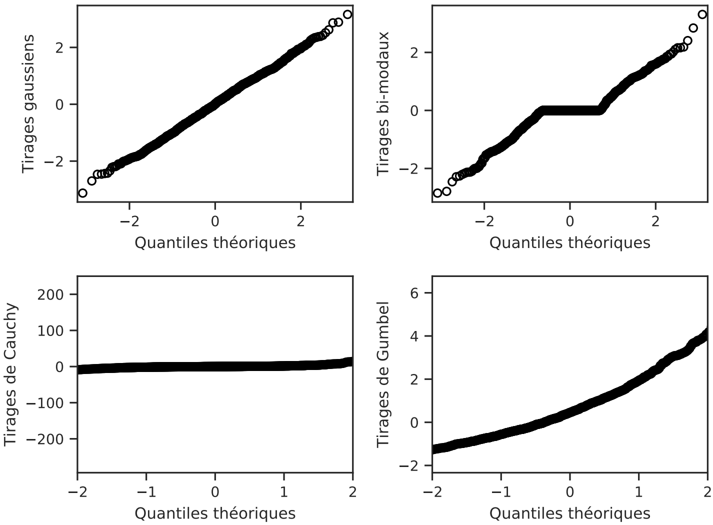
$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right), \text{ for } i = 1, \ldots, n$$

(voir détails en TD / TP)

Interprétation qq-plot poids à la naissance vs loi normale

- Si les observations étaient $\mathcal{N}(0,1)$ alors le nuage de points se concentrerait autour de la droite $y = x$
- Si le nuage de points se concentre autour d'une droite mais pas $y = x$, disons $y = ax + b$
  - Si $b \neq 0 \implies$ Translation
  - Si $a \neq 1 \implies$ Changement d'échelle

Quelques qq-plots pathologiques (vs. loi normale)



Bibliographie I

- Wilk, M. B. and R. Gnanadesikan. "Probability plotting methods for the analysis for the analysis of data". In: *Biometrika* 55.1 (1968), pp. 1–17.