Probability Distributions: Continuous

Introduction to Data Science Algorithms

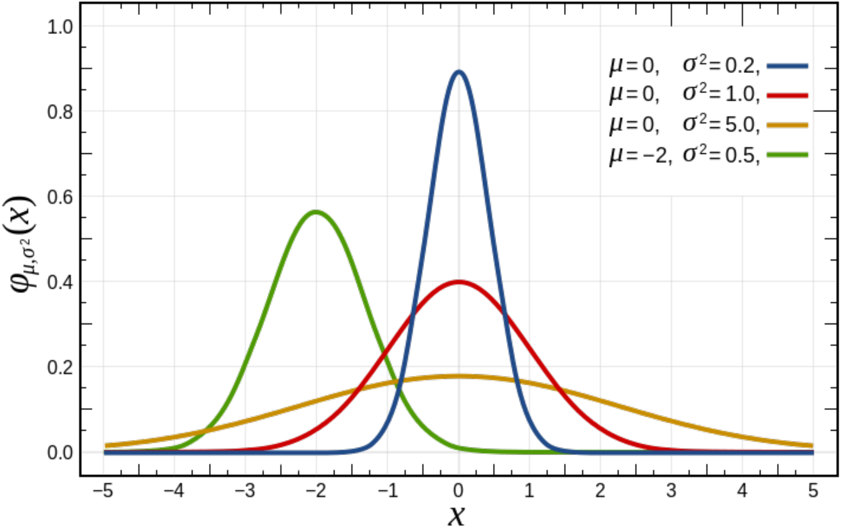
Dirk Grunwald

The normal distribution

- The most common continuous distribution is the *normal* distribution, also called the *Gaussian* distribution.
- The density is defined by two parameters:
  - μ: the *mean* of the distribution
  - σ^2: the *variance* of the distribution (σ is the *standard deviation*)
- The normal density has a “bell curve” shape and naturally occurs in many problems.

Carl Friedrich Gauss
1777 – 1855

The normal distribution



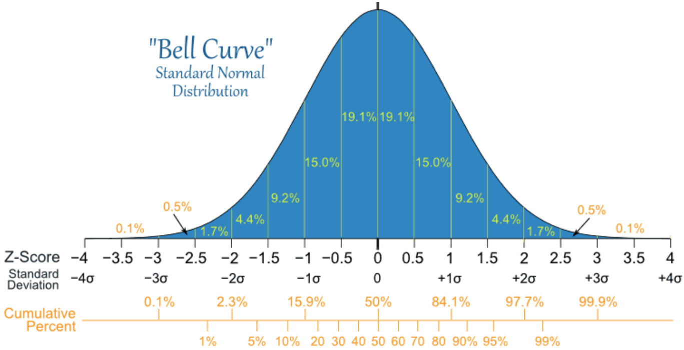
The normal distribution

- The probability density of the normal distribution is:

$$f(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\substack{\text{Does not} \\ \text{depend on } x}} \underbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\substack{\text{Largest when } x=\mu; \\ \text{shrinks as } x \text{ moves} \\ \text{away from } \mu}}$$

- Notation: $\exp(x) = e^x$
- If X follows a normal distribution, then $\mathbb{E}[X] = \mu$.
- The normal distribution is symmetric around μ.

The normal distribution



Standard Normal

- $Z \sim \mathcal{N}(0,1)$ is the *standard normal distribution*
- All normal distributions can be cast into *standard normal* using $X \sim N(\mu,\sigma)$ transformed into $Z = (X-\mu)/\sigma$

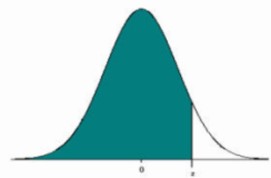
Standard Normal

- $Z \sim \mathcal{N}(0,1)$ is the *standard normal distribution*
- All normal distributions can be cast into *standard normal* using $X \sim N(\mu, \sigma)$ transformed into $Z = (X - \mu)/\sigma$
- Assume people are $\sim N(6, 0.1)$. What's the probability that someone is less than 6.05 feet tall?
    - Let $z = (6.05 - 6)/.1 = 0.4999$
    - Look up z in standard normal table and get 0.69
    - Thus, 69% of people are less than 6.05 feet tall (assuming $N(6, 0.1)$)

The Standard Normal Table

- Every stats book used to have standard normal table in the back.
- Thank goodness for computers!

Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

Applying the normal distribution

- Most variables in the real world don't follow an exact normal distribution, but it is a very good approximation in many cases.
  - Measurement error (e.g., from experiments) is often assumed to follow a normal distribution.
  - Biological characteristics (e.g., heights of people, blood pressure measurements) tend to be normal distributed.
  - Test scores – *e.g.* $IQ \sim N(100, 10)$

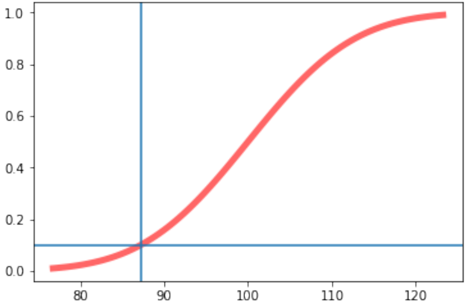
- Why? Central Limit Theorem
  - The *central limit theorem* proves that if you take the sum of multiple randomly generated values, the sums will follow a normal distribution. (Even if the randomly generated values do not!)

Quantiles

Let F be the CDF of a random variable X and let p be an arbitrary number between 0 and 1. The p^{th} quantile or $p \times 100^{th}$ percentile of the distribution of X is the smallest number q_p such that

$$F(q_p) = P(X \leq q_p) = p$$

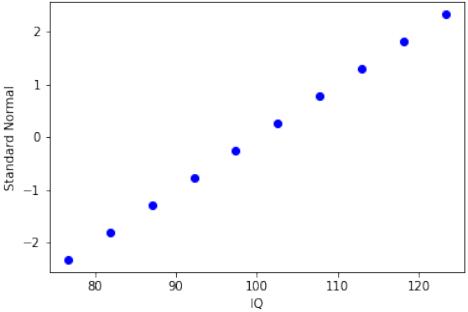
- Assume $IQ \sim N(100, 10) = F$.
- $F(87.18) = 0.10$, or $P(X \leq 87.18) = 0.10$
- This means $F^{-1}(0.10) = 87.18$
- The .10-quantile is 87.18



Quantile-Quantile Plots

If the "shape" of two distribution are similar, then a plot of the q^{th}-quantile of each distribution will be linear.

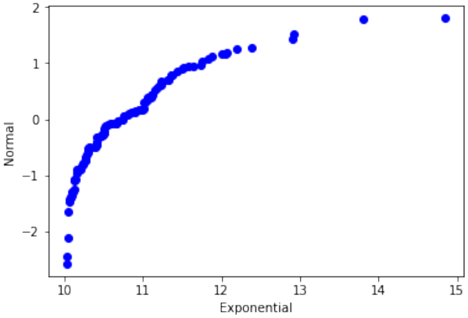
- Vertical axis $\sim Z$
- Horizontal axis $\sim N(100, 10)$
- In each case, we're showing the $0.1, 0.2, \ldots 0.8, 0.9$ quantile from easy distribution.
- Linear relationship indicates that test data (height) has the same "shape" as standard normal.
- Ergo, it's likely normal.



Quantile-Quantile Plots

If the plot of the q^{th}-quantile of each variable is not linear, than the variables are likely not from the same distribution.

- Vertical axis $\sim Z$
- Horizontal axis $\sim Exp(10)$
- We've draw 100 samples from each distributions, sorted them and then plotted the i^{th} sample in each case.
- non-linear relationship indicates that test data (Exp) does not have the "shape" as standard normal.
- Ergo, it's likely not normal.

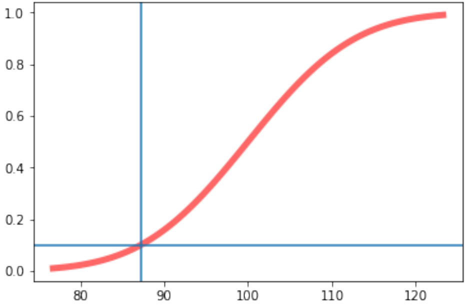


Quantiles

Let F be the CDF of a random variable X and let p be an arbitrary number between 0 and 1. The p^{th} quantile or $p \times 100^{th}$ percentile of the distribution of X is the smallest number q_p such that

$$F(q_p) = P(X \leq q_p) = p$$

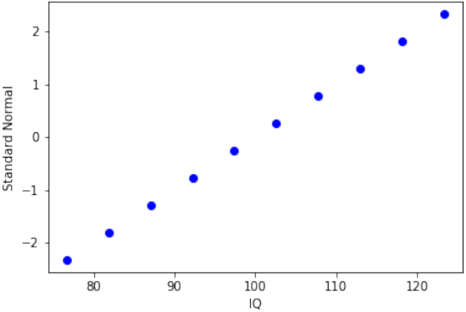
- Assume $IQ \sim N(100, 10) = F$.
- $F(87.18) = 0.10$, or $P(X \leq 87.18) = 0.10$
- This means $F^{-1}(0.10) = 87.18$
- The .10-quantile is 87.18



Quantile-Quantile Plots

If the “shape” of two distribution are similar, then a plot of the q^{th}-quantile of each distribution will be linear.

- Vertical axis $\sim Z$
- Horizontal axis $\sim N(100, 10)$
- In each case, we’re showing the $0.1, 0.2, \ldots 0.8, 0.9$ quantile from easy distribution.
- Linear relationship indicates that test data (height) has the same “shape” as standard normal.
- Ergo, it’s likely normal.



Quantile-Quantile Plots

If the plot of the q^{th}-quantile of each variable is not linear, than the variables are likely not from the same distribution.

- Vertical axis $\sim Z$
- Horizontal axis $\sim Exp(10)$
- We've draw 100 samples from each distributions, sorted them and then plotted the i^{th} sample in each case.
- non-linear relationship indicates that test data (Exp) does not have the "shape" as standard normal.
- Ergo, it's likely not normal.