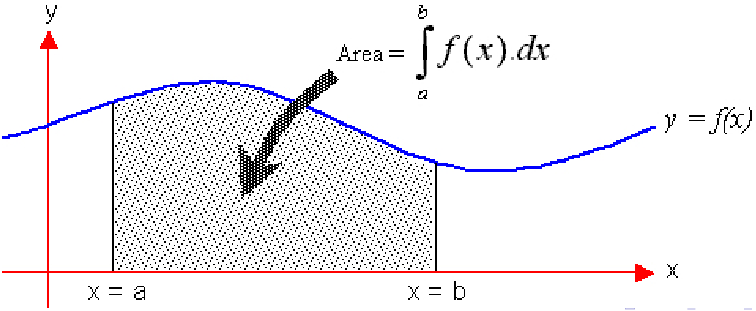
COMP229: Introduction to Data Science
Lecture 16: the normal distribution

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

Basics of integrals and areas

Definition 16.1. For a "good" function $f : \mathbb{R} \to \mathbb{R}$ (say, $f > 0$ for simplicity), the *integral* $\int_a^b f(x)dx$ is the area between the graph of the function $y = f(x)$ and the x-axis over the segment $[a, b]$.



A continuous random variable

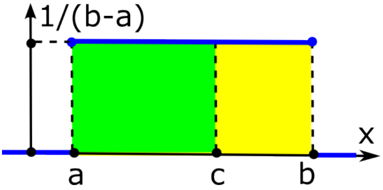
Definition 16.2. A continuous *random* variable X is given by a *probability density* $f(x) \geq 0$ such that the probability that X takes values smaller than a real number b is $P(X < b) = \int\limits_{-\infty}^{b} f(x)dx$.

If $f(x)$ quickly tends to 0 when $x \to \pm\infty$, the area over $(-\infty, b]$ can be computed as $\int\limits_{-\infty}^{b} f(x)dx$.

When $b \to +\infty$, the probability $P(X < b)$ tends to 1, so any probability density has $\int\limits_{-\infty}^{+\infty} f(x)dx = 1$.

A uniform random variable

Definition 16.3. A *uniform* variable over $[a, b]$ has the density $f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a, b], \\ 0 & \text{for } x \notin [a, b]. \end{cases}$



The rectangle with sides $b - a$ and $\dfrac{1}{b-a}$ has the area $\int_{-\infty}^{+\infty} f(x)dx = 1$.

For $c \in [a, b]$, the probability $P(X < c)$ equals $\dfrac{c-a}{b-a} =$ the area of the rectangle $[a, c] \times [0, \dfrac{1}{b-a}]$.

A normal random variable

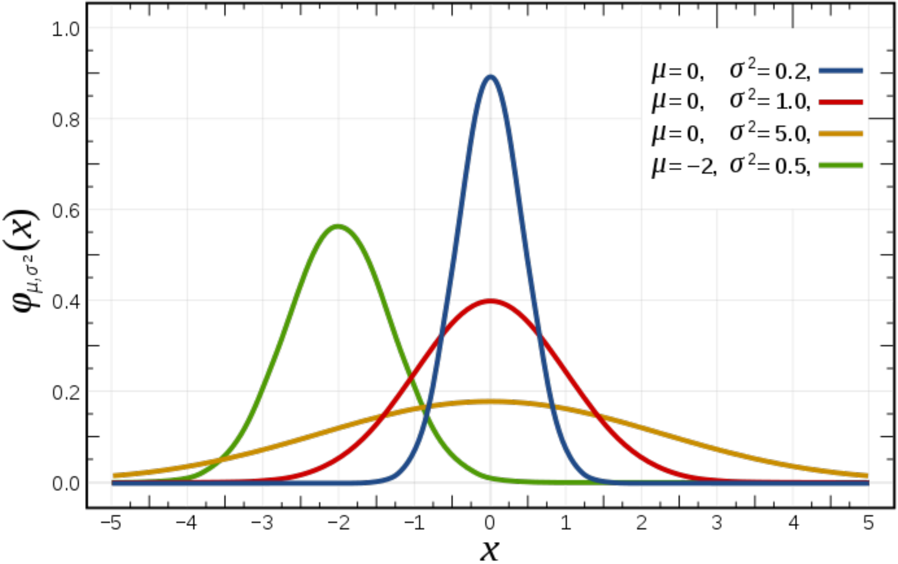
Definition 16.4. A *normal variable* X has the density $\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, where μ is the mean (also the median and mode), σ is the standard deviation. X can be shortly introduced as $X \sim N(\mu, \sigma^2)$, where σ^2 is the *variance*.

The standard normal variable is $X \sim N(0, 1)$.

The factor $\frac{1}{\sqrt{2\pi}\sigma}$ implies that $\int\limits_{-\infty}^{+\infty} \phi_{\mu,\sigma^2}(x)dx = 1$.

$\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \to 0$ quickly when $x \to \pm\infty$.

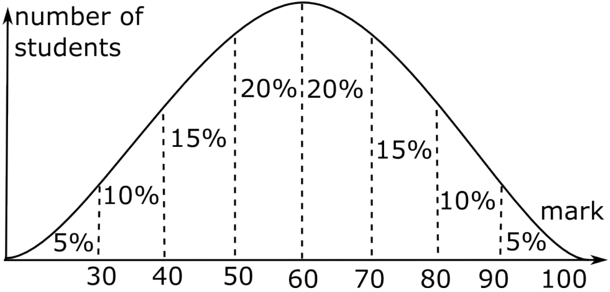
Normal densities for various μ, σ



The central limit theorem

Claim 16.5. If $X_1, \ldots, X_n$ are independent identically distributed variables with variance σ^2, mean 0, then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to N(0, \sigma^2)$ as $n \to +\infty$.

Informally, "in the limit any average is normal".



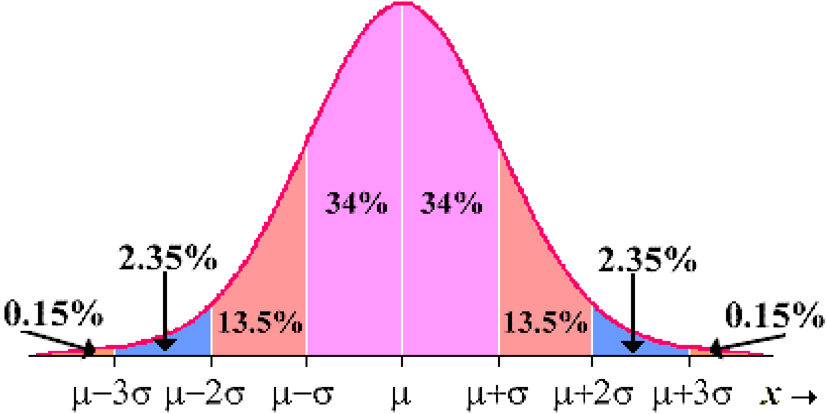
One often assumes that random variables that we can't control are normal.

The 68-95-99.7 (approximate) rule

About 68% of observations are in $[\mu - \sigma, \mu + \sigma]$.

About 95% of observations are in $[\mu - 2\sigma, \mu + 2\sigma]$.

About 99.7% of observations are in $[\mu - 3\sigma, \mu + 3\sigma]$.



The standardized Z-score

Claim 16.6. If a normal variable X has a mean μ and a standard deviation σ, the *standardized score* $Z = \dfrac{X-\mu}{\sigma}$ has the density $N(0,1)$.

Problem. Let exam marks have a normal distribution with $\mu = 60$ and $\sigma = 10$. What proportion of the class failed the exam?

Solution. The required proportion is the probability $P(X < 40) = P(X < \mu - 2\sigma) = 2.5\%$.

If it's hard to express 40 via μ, σ, use the Z-score.

Using the Z-score

The bound 40 of the given random variable $X \sim N(60, 10^2)$, for the Z-score $Z = \dfrac{X - 60}{10}$ becomes $\dfrac{40 - 60}{10} = -2$, so we need $P(Z < -2)$ for the standard normal variable $Z \sim N(0, 1)$.

The probability $P(Z < -2)$ is 2.5% from the 68–95–99.7 rule. The proportion of the students who passed is $P(X > 40) = P(Z > -2) = 97.5\%$

Find the proportion of students with 70+ marks.

Your questions and the quiz

$P(X > 70) = P(Z > 1) = 16\%$ by the 68% rule.

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Let $X \sim N(60, 20^2)$. Find $P(X < 40)$.

Answer to the quiz and summary

Answer. The bound 40 for X becomes -1 for $Z = \dfrac{X-60}{20}$, so $P(X < 40) = P(Z < -1) = 16\%$.

- The *normal* random variables has the probability density $\dfrac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$, where μ is the mean, σ is the standard deviation
- For $X \sim N(\mu, \sigma^2)$, the *standardized* variable is $Z = \dfrac{X-\mu}{\sigma}$ whose probabilities $P(Z < b)$ are pre-computed (in available tables online).