COMP229: Introduction to Data Science
Lecture 9: When Everything is Normal

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Recap: continuous random variables

- A random variable X is called **continuous** if its CDF F_X is a continuous function.
- X can also be defined by a PDF $f(x) \geqslant 0$.
- PDF and CFD can be defined from each other: $P(X < v) = F_X(v) = \int_{-\infty}^{v} f(x)dx$ and $f(x) = \dfrac{dF(x)}{dx}$.

A normal random variable

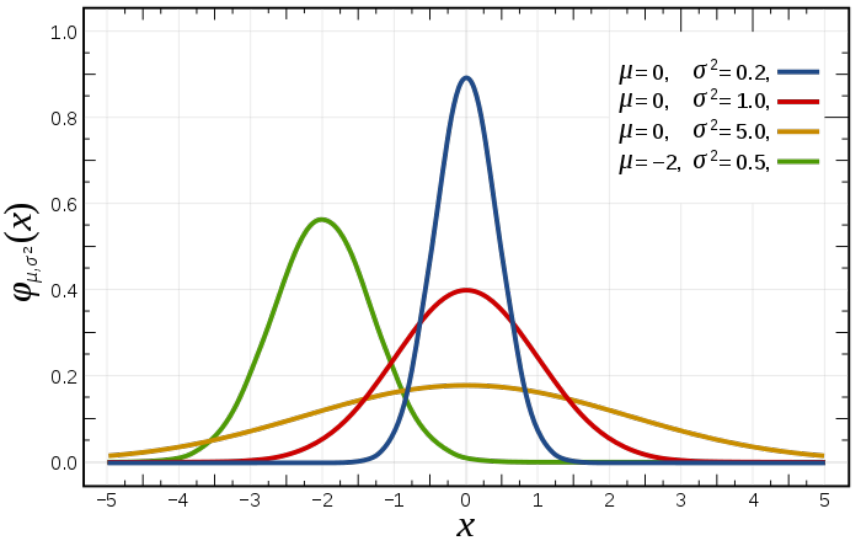
Definition 9.1. A **normal variable** $X \sim N(\mu, \sigma^2)$ has the density $\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$,
where μ is the *mean* (also the median and mode), σ is the standard *deviation*.

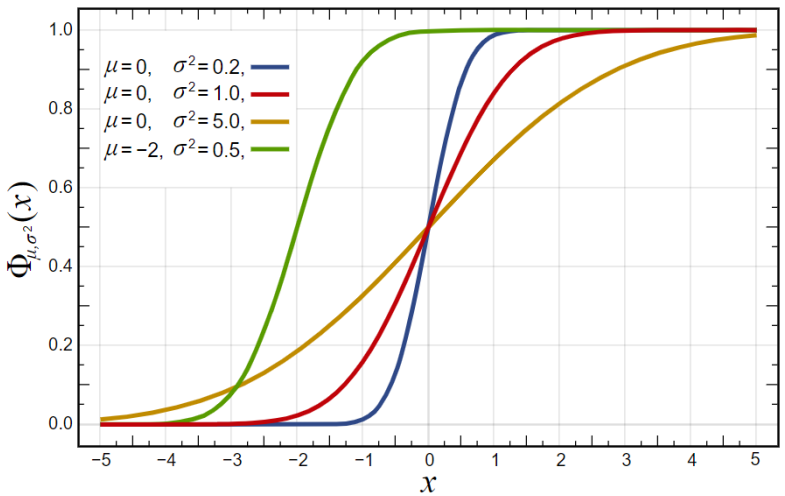
The **standard normal** variable is $X \sim N(0, 1)$.

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

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

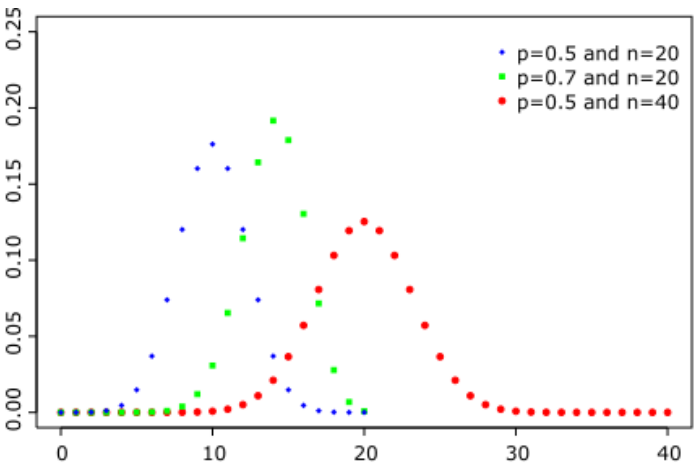
Normal distributions for various μ, σ

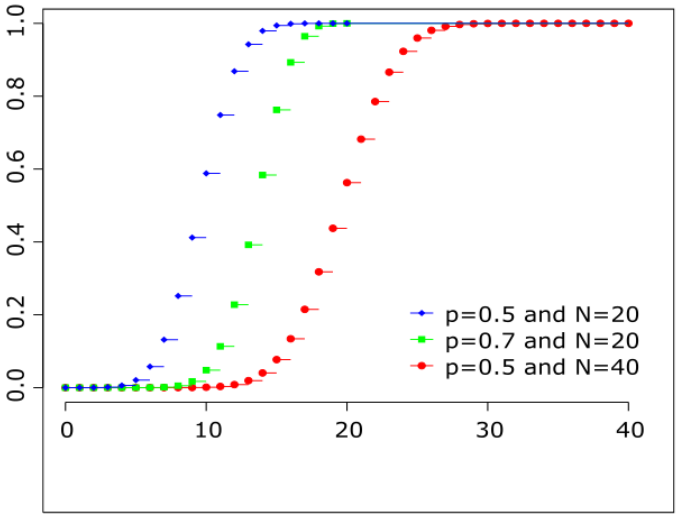




Reminder: Binomial distribution

Binomial discrete distribution $X \sim B(n, p)$, which is the number of successes in a sequence of n independent Bernoully (binary success/failure) distributions with probability p of success, looked similar:



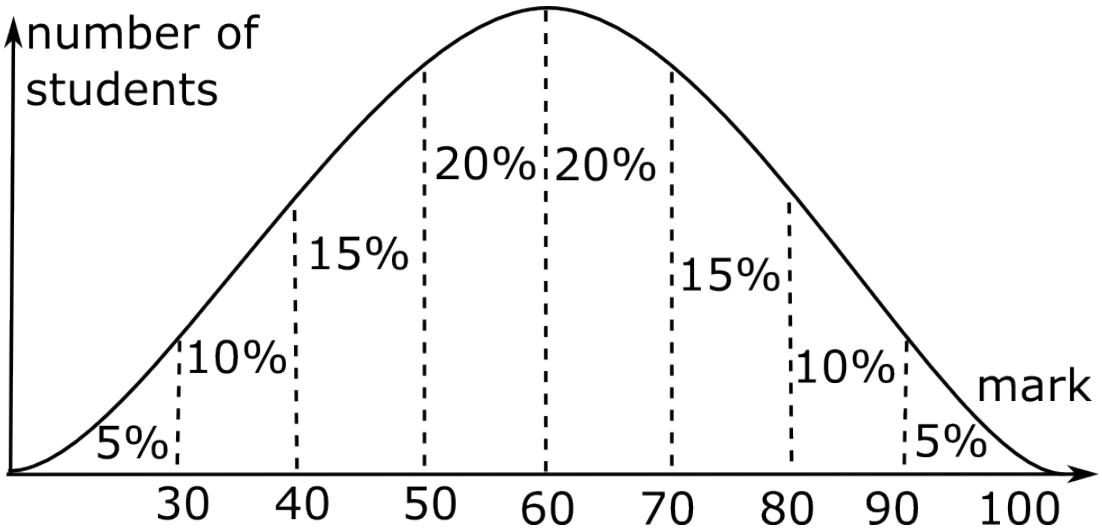


Is it just a coincidence? Try this experiment.

The Central Limit Theorem

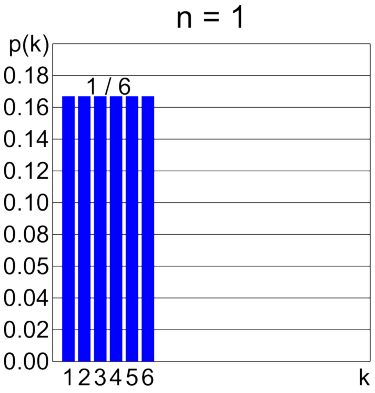
Theorem 9.2. If $X_1, \ldots, X_n$ are *independent identically distributed (i.i.d.)* variables with variance σ^2 and mean 0, then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \underset{n \to +\infty}{\to} N(0, \sigma^2)$.

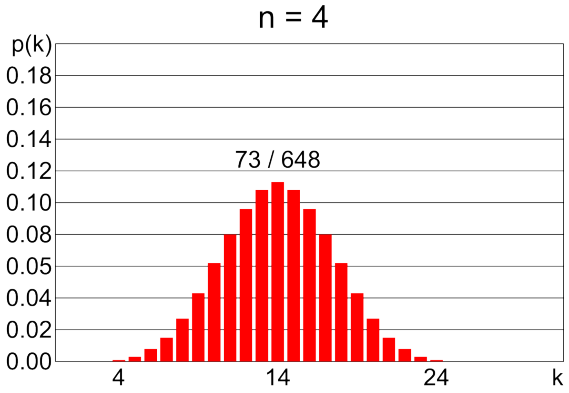
Informally, "in the limit any average is normal".

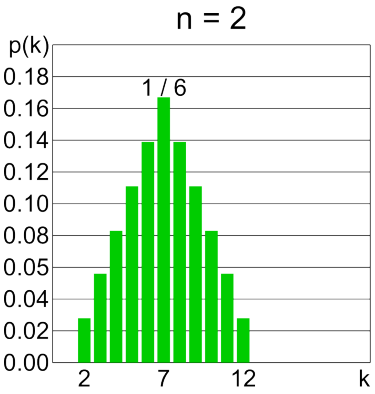


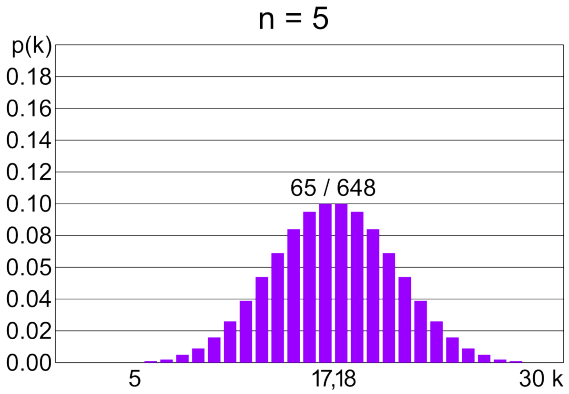
One often assumes that random variables that we can't control are normal.

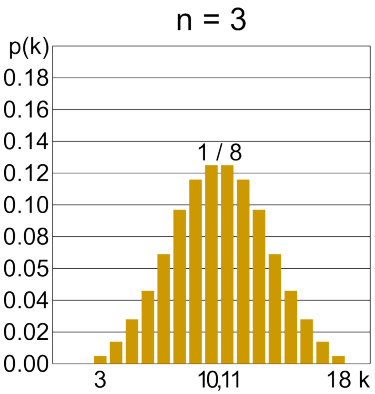
CLT for dice











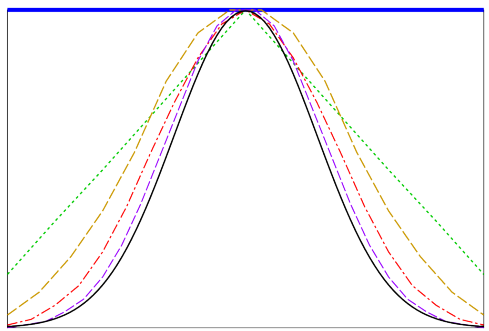


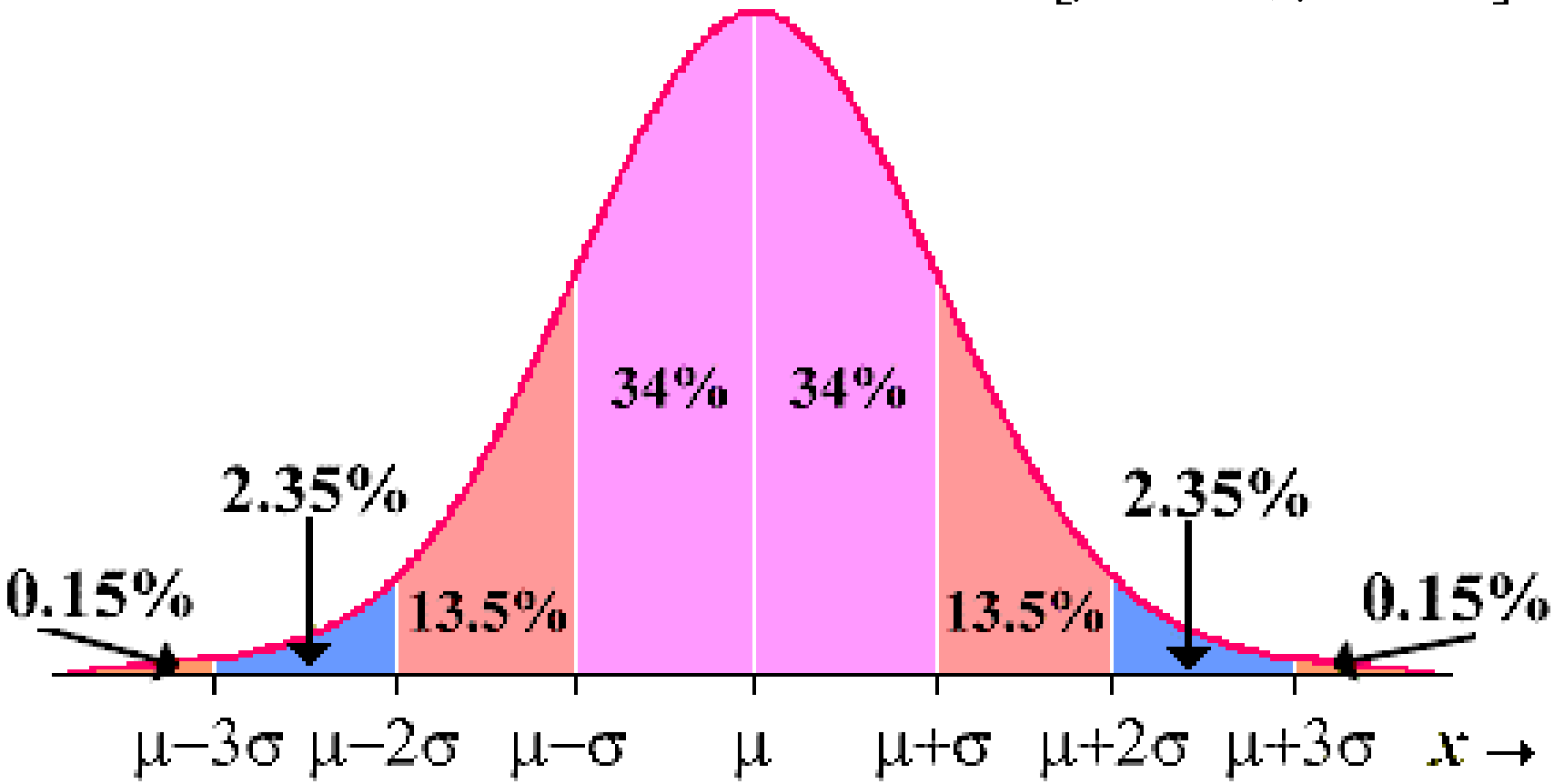
Image by Cmglee - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18918612

The $\sigma - 2\sigma - 3\sigma$ 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 9.3. If $X \sim N(\mu, \sigma^2)$, then the *standardized score* $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ has the density $\dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$.
Then $P(X < v) = P(\frac{X-\mu}{\sigma} < \frac{v-\mu}{\sigma}) = P(Z < \frac{v-\mu}{\sigma})$.

Problem 9.4. Let exam marks have a normal distribution with $\mu = 60$ and $\sigma = 10$.
1) What proportion of the class has failed the exam?
2) Find the proportion of 70+ marks.

Solution 9.4. 1) The required proportion is the probability $P(X < 40) =$

The standardized Z-score

Claim 9.3. If $X \sim N(\mu, \sigma^2)$, then the *standardized score* $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ has the density $\dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$.
Then $P(X < v) = P(\frac{X-\mu}{\sigma} < \frac{v-\mu}{\sigma}) = P(Z < \frac{v-\mu}{\sigma})$.

Problem 9.4. Let exam marks have a normal distribution with $\mu = 60$ and $\sigma = 10$.
1) What proportion of the class has failed the exam?
2) Find the proportion of 70+ marks.

Solution 9.4. 1) The required proportion is the probability $P(X < 40) = P(X < \mu - 2\sigma) \approx 2.5\%$.

Standardising into Z-score

If it's hard to express 40 via μ, σ, use 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 $Z \sim N(0, 1)$.

Now it's easy to use the $\sigma - 2\sigma - 3\sigma$ rule, $P(Z < -2) \approx 2.5\%$.

2) By the 68% rule $P(X > 70) =$

Standardising into Z-score

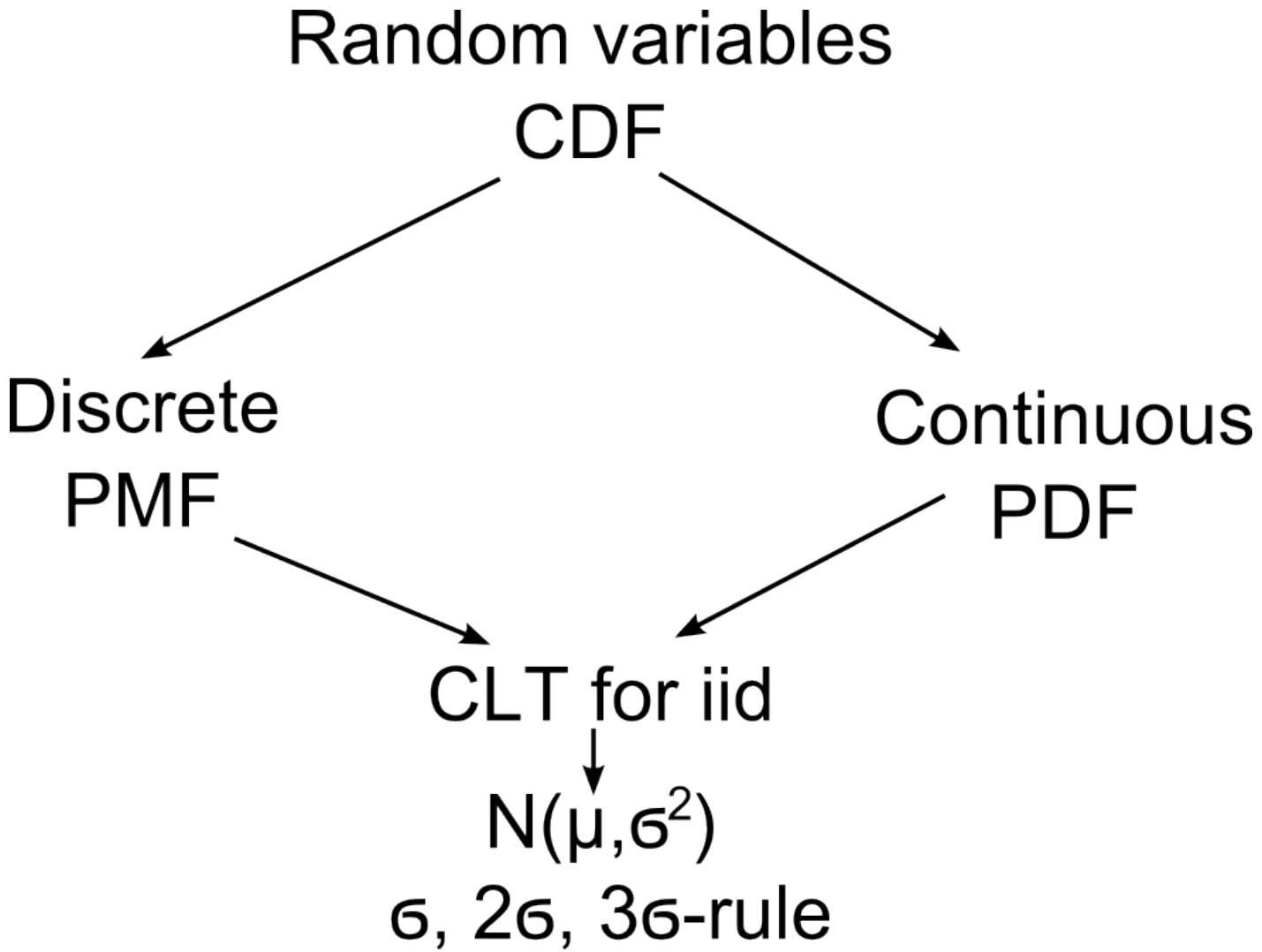
If it's hard to express 40 via μ, σ, use the Z-score.

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 $Z \sim N(0, 1)$.

Now it's easy to use the $\sigma - 2\sigma - 3\sigma$ rule, $P(Z < -2) \approx 2.5\%$.

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

Connecting discrete and continuous



Statistical inference

Probability theory studies problems when a probability distribution is (assumed to be) known, e.g. the normal probability density with a given mean μ and a standard deviation σ.

In practice we have sample data and wish to make conclusions about a population, e.g. about parameters (μ and σ) of a normal density.

Drawing such conclusions is *statistical inference*.

Point estimates vs interval estimates

A *point* estimate is a single value estimated from a sample, e.g. the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a point estimate of the mean of a normal density.

An *interval* estimate specifies an interval containing the estimated parameter, e.g. the mean μ is within $[\bar{x} - \text{margin of error}, \bar{x} + \text{margin of error}]$. The last conclusion will come with a confidence, e.g. 95%.

Under what conditions can we do this?

Conditions for inference

Conditions or assumptions are a *statistical model*.

1. We have a simple random sample (SRS) from the larger population (theoretically infinite).

2. The variable we are interested in has the normal distribution $N(\mu, \sigma^2)$ in the very large population (not in the much smaller sample).

3. We know the standard deviation σ of the variable in question, but not the mean μ.

To make an estimate, we need sample values.

A typical problem to estimate μ

Assume that we know only 9 exam marks
$\{x_1, \ldots, x_9\} = \{60, 70, 80, 50, 90, 40, 30, 45, 75\}$
which are measurements of random variables
$X_1, \ldots, X_9 \sim N(\mu, \sigma^2)$.
Given $\sigma = 15$, estimate an interval for the mean μ with confidence 95%.

Step 1. Compute the sample mean $\bar{x} =$

$$\frac{60 + 70 + 80 + 50 + 90 + 40 + 30 + 45 + 75}{9} = 60.$$

The deviation of the average

Step 2. Considering $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as a random variable, estimate its standard deviation.

Claim 9.5. [no proof needed, follows from CLT] If variables $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has the normal distribution $N(\mu, \sigma^2/n)$.

The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$.

Estimating a probability

Step 3. For the required 95% confidence, the 68-95-99.7 rule says that the random variable $\bar{X}$ takes values within $2\sigma(\bar{X}) = 10$ of the mean μ with the probability 95%, i.e. $P(|\bar{X} - \mu| < 10) \approx 0.95$.

We'll rewrite the above conclusion for μ in terms of the sample average $\bar{x}$ and an error margin for

a required confidence $=$ probability of the event $\bar{X} \in [\bar{x} - \text{margin of error}, \bar{x} + \text{margin of error}]$.

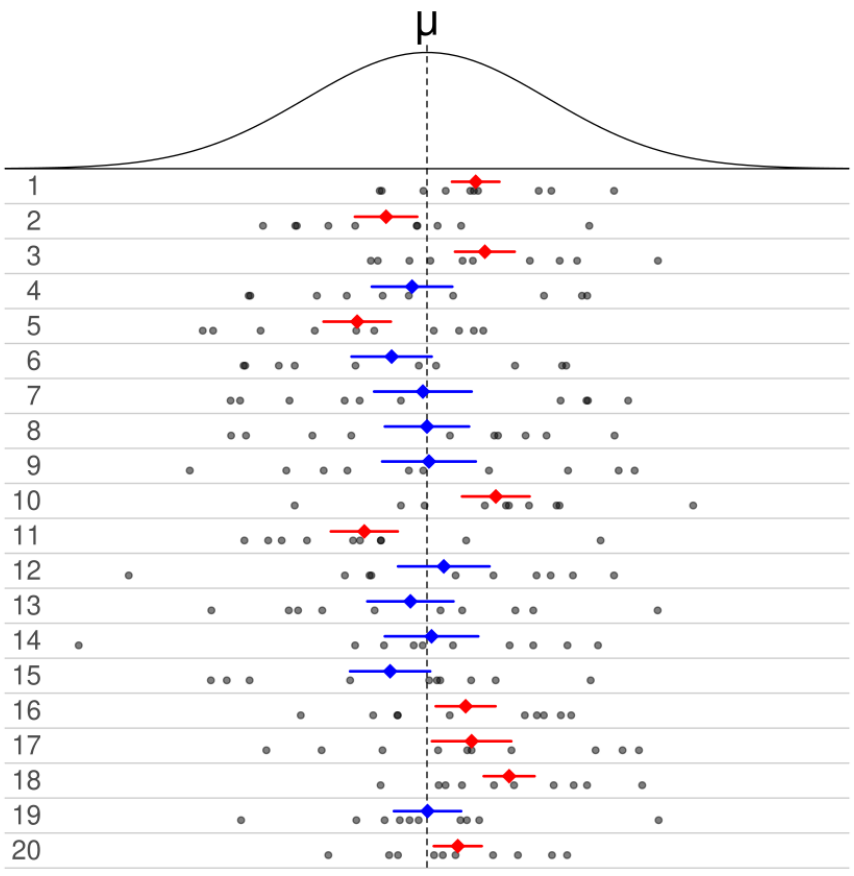
Obtaining the error margin

Step 4. $P(|\bar{X} - \mu| < 10) = 0.95$ means that for 95% of all samples μ can be found within ± 10 from the mean $\bar{x} = 60$. Hence we are 95% "confident" that μ could be within 60 ± 10.

The value after $\pm$ is the *margin of the error*.

Definition 9.6. Were this procedure to be repeated many times, the computed **confidence interval** will cover the true μ with probability (or **confidence level**) α.

"Catching" the parameter



from wikipedia.

Sample from the same normal distribution with the 50% confidence intervals for μ. The blue intervals contain the true μ, the red ones do not.

General critical values z^*

For a given probability P and normal variable $Z \sim N(0,1)$, the equation $P(|Z| < z^*) = P$ has numerical solutions in a "standard normal table".

The approximate 68-95-99.7 rule said that $P(|Z| < 2) \approx 0.95$, but the better value is 1.96.

confidence level	90%	95%	99%
critical value z^*	$1.645 \approx 1.7$	$1.96 \approx 2$	$2.576 \approx 2.6$

You could remember these values for the exam.

General case of step 3

Claim 9.7. Let z^* be the critical value satisfying $P(|Z| < z^*) = P$ for a given confidence level P and a normal variable $Z \sim N(0, 1)$. Assume that n samples are independently drawn from a normal distribution whose standard deviation is σ. Then the mean μ has the *confidence interval* $\bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$.

Problem 9.8. The sample 60, 70, 65, 80, 85, 35, 55, 50, 100, 20, 90, 60, 40, 30, 45, 75 is from a normal distribution with the standard deviation $\sigma = 20$. Estimate the mean with confidence 95%.

One more typical solution

Solution 9.8. The 16 values have the average

$$\bar{x} = (60 + 70 + 65 + 80 + 85 + 35 + 55 + 50 + 100 + 20 + 90 + 60 + 40 + 30 + 45 + 75)/16 = 60.$$

For confidence 95%, the critical value is $z^* = 1.96$.
The margin of error is $z^* \dfrac{\sigma}{\sqrt{n}} = 1.96 \times \dfrac{20}{\sqrt{16}} = 9.8$
and the mean μ is estimated between 60 ± 9.8.

μ is between 60 ± 12.88 with confidence 99%.

μ is between 60 ± 8.225 with confidence 90%.

Famous citation

All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true.

All models are wrong, but some models are useful.

So the question you need to ask is not "Is the model true?" (it never is) but "Is the model good enough for this particular application?"

George E.P.Box

Time to revise and ask questions

To estimate the mean μ for a confidence P using a known deviation σ and a sample of n values, do

- find the simple average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
- compute the standard deviation $\frac{\sigma}{\sqrt{n}}$ of $\bar{X}$;
- find the value z^* from $P(|Z| < z^*) = P$;
- write the interval for μ between $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.

Problem 9.9. For a normal variable $Z \sim N(0, 1)$, what is z^* satisfying $P(|Z| < z^*) = 0.99$?