3a - Probability, The Central Limit Theorem, Confidence Intervals

Probability

Probability = relative frequency of occurence of event at hand

- Probabilities = numbers assigned to sets
- these sets $\in$ universe Ω

Axioms of Probability

Rule	Formula
Probabilities are non-negative	$\forall A \in \Omega : \mathcal{P}(A) \geq 0$
Universe has probability 1	$\mathcal{P}(\Omega) = 1$
Sum rule for disjoint events	$A \cap B = \varnothing \implies \mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B)$

Properties of probabilities

Name	Formula
Complement rule	$\mathcal{P}(\overline{A}) = 1 - \mathcal{P}(A)$
Impossible event	$\mathcal{P}(\varnothing) = 0$
General sum rule	$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$

Independent event

$$A \perp\!\!\!\perp B \iff \mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B)$$

$\implies$ for dependent events, $=$ changes to $\neq$

Discrete random variables

Discrete = finite set of possible outcomes

Probability mass function

The probability that variable X is value x.

$$f_X(x) = \mathcal{P}(X = x) =$$

You can use it to compute other probabilities, e.g. $\mathcal{P}(X \leq 2) = \mathcal{P}(X = 0) + \mathcal{P}(X = 1) + \mathcal{P}(X = 2)$

Expectation and variance of a random variable

Expectation of a random variable

$$\mu_X = \sum_i x_i \mathcal{P}(X = x_i) = \sum_i x_i f_X(x_i)$$

Also written as μ

Variance of a random variable

$$\sigma_X^2 = \sum_i (x_i - \mu_X)^2 \mathcal{P}(X = x_i) = \sum_i (x_i - \mu_X)^2 f_X(x_i)$$

Standard deviation = positive sqrt of variance:

$$\sigma_X = \sqrt{\sigma_X^2}$$

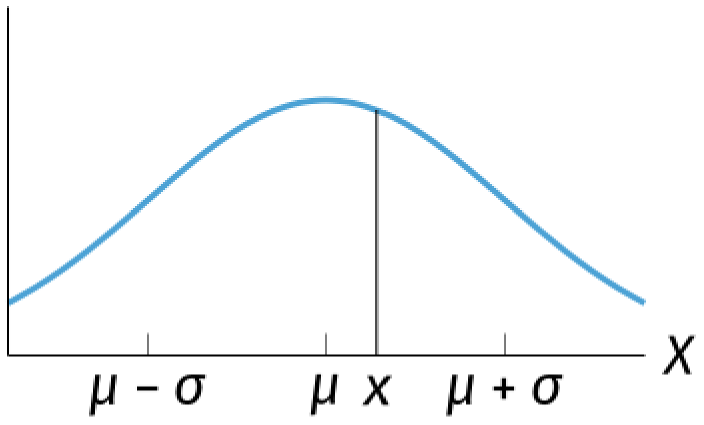
Continuous Random Variables

- Continuous = infinite possibilities
- Probability that a is *exactly* something is always 0
  $\implies$ consider intervals
- probability can be found by *integrating* the probability density function

The (Standard) Normal Distribution

Normal distribution

$$x \in X \sim \mathrm{Nor}(\mu, \sigma)$$



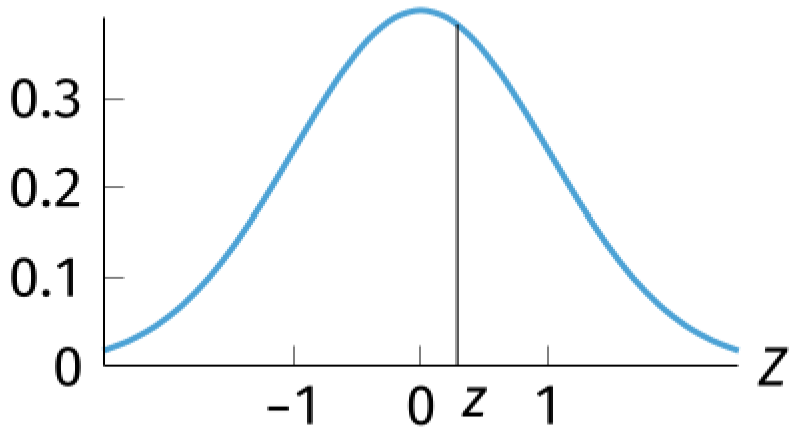
- Expectation

$$\mu_X = \int_{-\infty}^{+\infty} x f_X(x) dx$$

- Variance

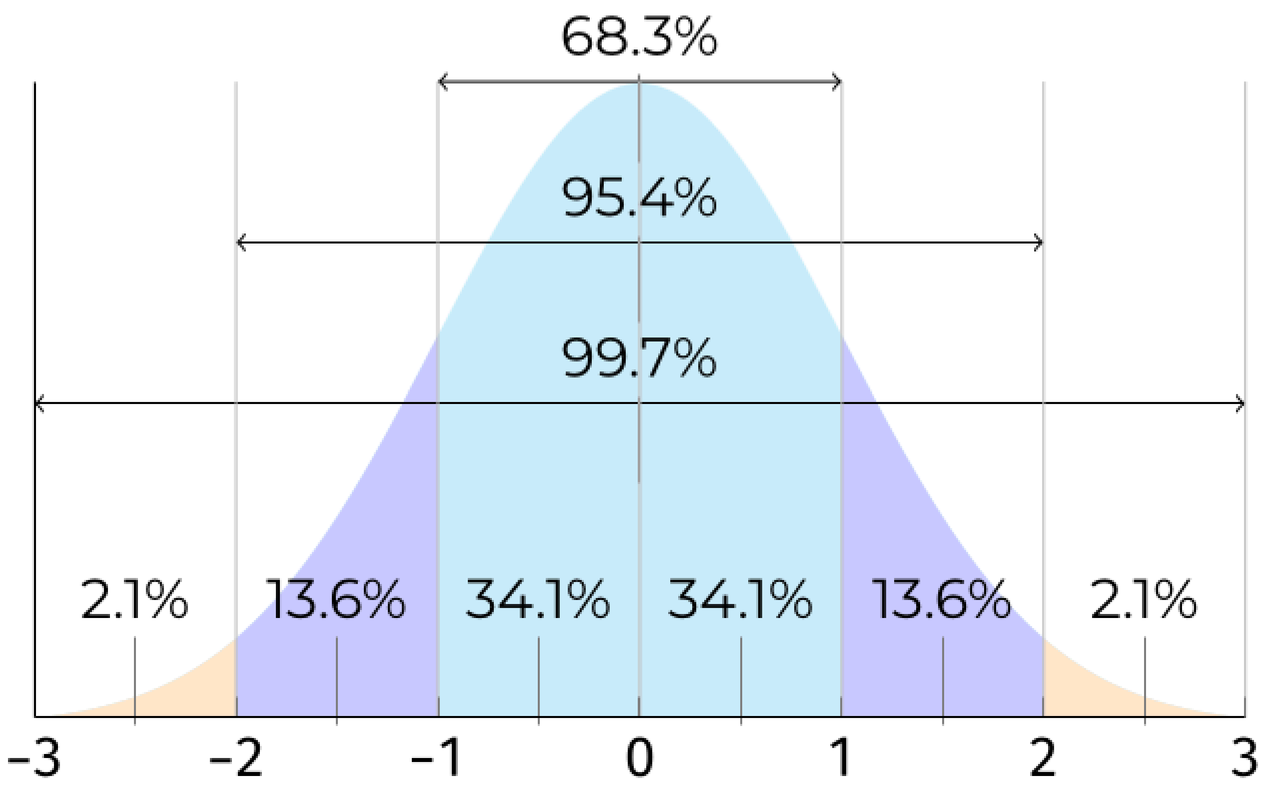
$$\sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x) dx$$

Standard normal distribution



$$x \in X \sim \mathrm{Nor}(0,1)$$

- x and z have similar position in Gaussian bell curve
  $\implies x = \mu + z \cdot \sigma$ and $z = \frac{x-\mu}{\sigma}$



Probabilities in the Normal Distribution

Python functions

`import scipy.stats`

Function	Prupose
`norm.pdf(c, loc=m, scale=s`	Probability density at `x`
`norm.cdf(x, loc=m, scale=s)`	Left-tail prbability $\mathcal{P}(X < x)$
`norm.sf(x, loc=m, scale=s)`	Right-tail probability $\mathcal{P}(X > x)$
`norm.isf(1-p, loc=m, sclale=s)`	$p\%$ of observations are expected to be lower than result

e.g. $X \sim \mathrm{Nor}(\mu = 5, \sigma = 1.5) : \mathcal{P}(X > 6) =$ `stats.norm.sf(6, loc=5, scale=1.5)`

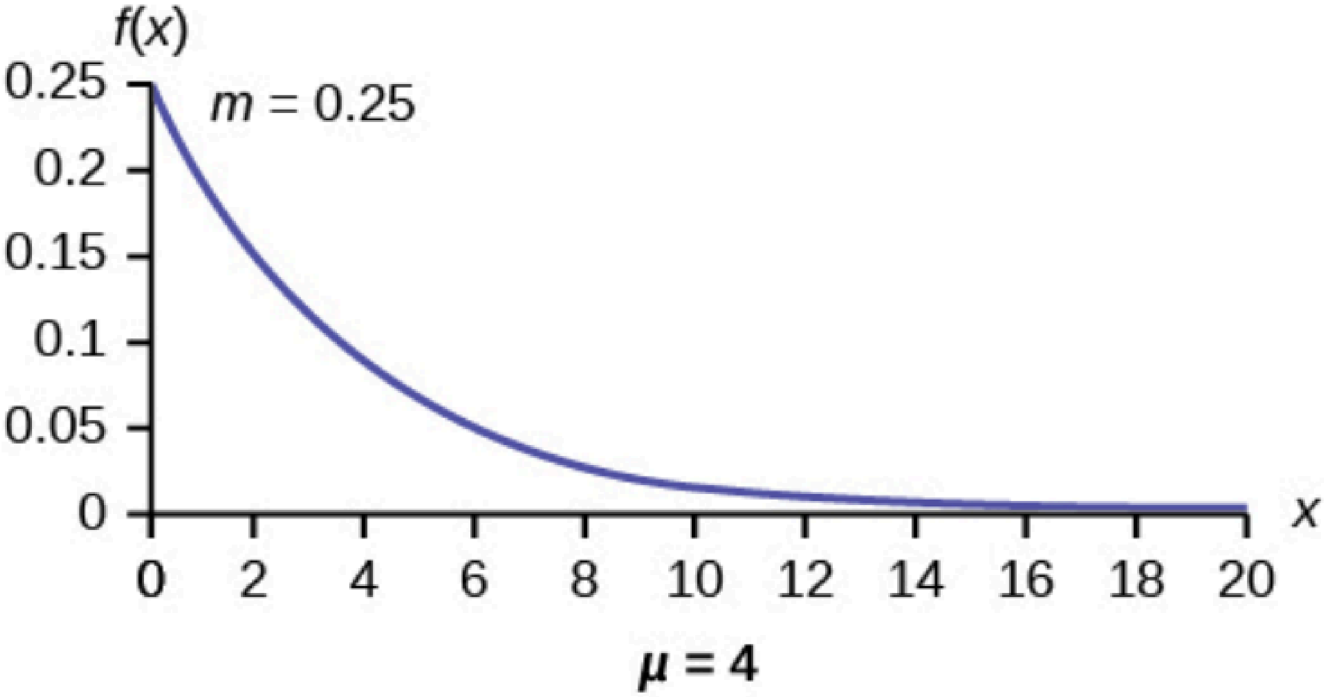
Other Continuous Distributions

Exponential distribution

when more smaller values than larger values
e.g.

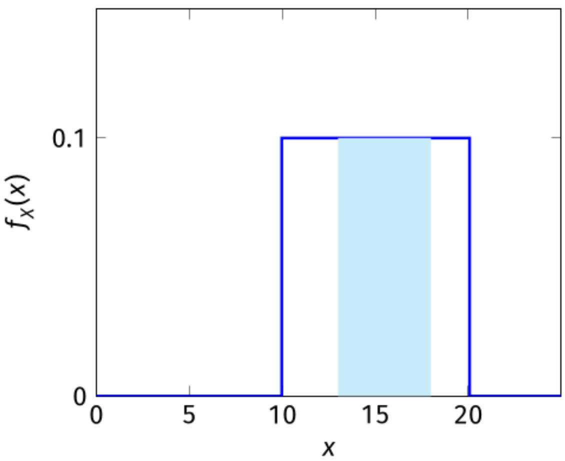
- money spent in grocery store per customer

- amount of time clerk spends with their customers



Continuous uniform distribution

- has arbitrary outcome between two bounds
- density function = constant



$P(13 < X < 18)$ = area under $f_X(x)$ between 13 & 18
$= 5 \times 0.1$
$= 0.5$

Note:

$P(13 < X < 18) = P(13 < X \leq 18) = P(13 \leq X < 18)$
$= P(13 \leq X \leq 18)$
and it also holds that

$$P(10 \leq X \leq 20) = 1$$

so the total area under the probability density function is always equal to 1.

From Sample to Population

The Central Limit Theorem

If the size is sufficiently large, the **probability distribution** of the **sample mean** will approximate a normal distribution regardless of probability distribution of underlying population.

sample of n observations from population with expected value μ and standard deviation σ. If n sufficiently large:
$\overline{x} \sim \text{Nor}(\mu = \mu_{\overline{x}}, \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}})$

Point estimate

point estimate for population parameter = formula / equation to calculate value estimate for that parameter. (e.g. sample variance & standard deviation)

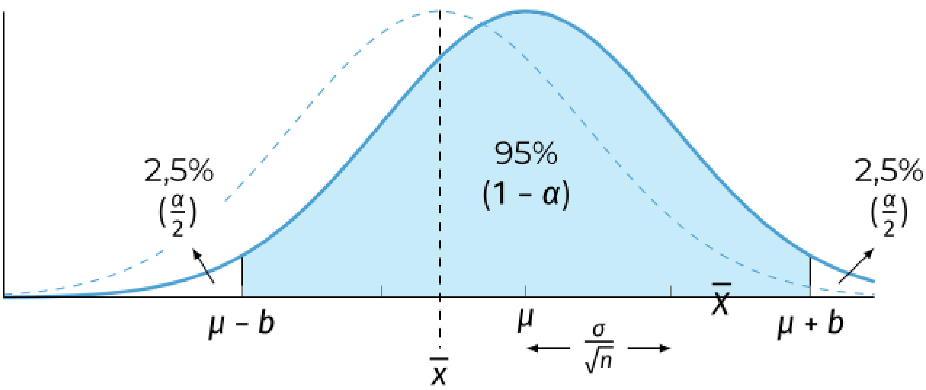
Confidence Intervals

Confidence interval = equation / formula to construct interval that will contain parameter to be estimated with certain leverl of confidence.

Confidence Interval for a Large Sample

Because of the central limit theorem we know that: $\bar{x} \in \bar{X} \sim Nor(\mu, \frac{\sigma}{\sqrt{n}})$
And because of the symmetry we can say:

$$P\left(\bar{x} - b < \mu < \bar{x} + b\right) = P\left(\mu - b < \bar{x} < \mu + b\right)$$

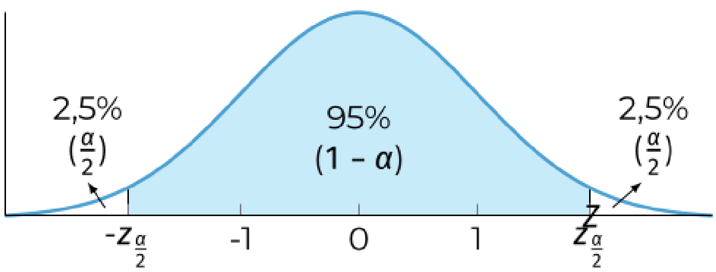


We now calculate the z-score for $\bar{x}$: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
We lookup (or calculate) the value $z_{\frac{\alpha}{2}}$ for which:

$$P\left(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha = 0.95$$

$$P\left(z < z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} = 0.975$$

$$z_{\frac{\alpha}{2}} = \texttt{stats.norm.isf(1-0.975)} \approx 1.96$$



The result is:

$$P\left(-1,96 < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1,96\right) = 0,95$$

$$P\left(\mu - 1,96\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Because of symmetry we can swap μ and $\overline{x}$:

$$P\left(\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Now we can say with 95% confidence that:

$$\mu \in \left[\ \overline{x} - 1,96.\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + 1,96.\frac{\sigma}{\sqrt{n}}\ \right]$$

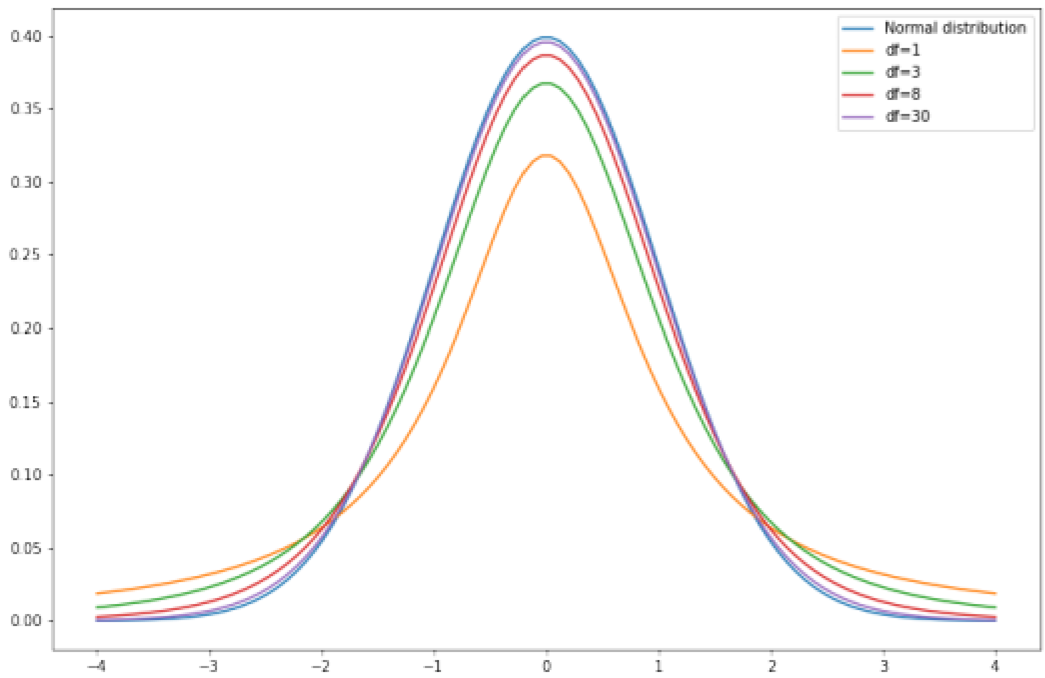
(in practice we will use s_{sample} as a point estimate for $\sigma_{population}$)

Confidence Interval for a Small Sample

if $X \sim \mathrm{Nor}(\mu, \sigma)$ and you take a small sample with mean $\overline{x}$ and stddev s, then

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

will behave as t-distribution with $n - 1$ deg of freedom



t-distrbution in Python

df = degrees of freedom

Functie	Betekenis
`t.pdf(x, df=d)`	Probability density for x
`t.cdf(x, df=d)`	Left-tail probability $P(X < x)$
`t.sf(x, df=d)`	Right-tail probability $P(X > x)$
`t.isf(1-p, df=d)`	p% of observations is expected to be lower than this value

Confidence interval

To determine the confidence interval for the mean μ of a population, based on a *small* sample, we first calculate $t_{\frac{\alpha}{2},df}$.

With a confidence level of 95% we have $\frac{\alpha}{2} = 0,025$
Assume for example $n = 5$ (so `df=4`), then we have

$$t_{\frac{\alpha}{2},df} = \texttt{stats.t.isf(1-0.975, df=4)} = 2.776$$

We can say with a confidence of 95% that:

$$\mu \in \left[\overline{x} - t_{\frac{\alpha}{2},df} \cdot \frac{s}{\sqrt{n}} \;,\; \overline{x} + t_{\frac{\alpha}{2},df} \cdot \frac{s}{\sqrt{n}} \right]$$