Topic 4 – Normal and sampling distributions

ENVX1002 Introduction to Statistical Methods

Dr. Floris van Ogtrop
The University of Sydney

Dec 2024

THE UNIVERSITY OF SYDNEY

Topic 4 Outline – Normal and sampling distributions

- Example
- Normal distribution
- Other continuous distributions
- Sampling distribution

Learning Outcomes

- Understand what a (probability) distribution is:
  - the properties of a continuous distribution.
- Use Normal Distribution to understand/describe data
  - Be able to standardise a Normal;
  - Calculate probabilities based on Normal Distribution using R.
- Know that there are other continuous distributions useful in hypothesis testing.
- Distinguish between population, sample and sampling distributions;
- Distinguish between a standard deviation and standard error of the mean;
- Describe the Central Limit Theorem;
- Use R and Excel to calculate the standard error and probabilities associated with sampling distributions;

Types of data

- Numerical
  - Continuous: yield, weight
  - Discrete: weeds per m^2
- Categorical
  - Binary: 2 mutually exclusive categories
  - Ordinal: categories ranked in order
  - Nominal: qualitative data

Example

- The gestation period (in days) for American Simmental cattle is distributed with mean 284.3 and standard deviation 5.52. How often is a calf born a week early?

Wray et al. 1987

J Anim Sci. 1987 Oct;65(4):970-4. doi: 10.2527/jas1987.654970x.

Analysis of gestation length in American Simmental cattle

N R Wray [1], R L Quaas, E J Pollak

Affiliations + expand

PMID: 3667470 DOI: 10.2527/jas1987.654970x

Abstract

Records of gestation length (71,461) for Simmental cattle were distributed with mean 284.3 d and standard deviation 5.52 d. Gestation length was found to increase with percent Simmental and was 1.9 d longer for calves born to mature dams than for those born to heifer dams. Bull calves experienced gestation lengths 1.5 d longer than heifer calves. Sire, maternal grandsire, residual and total variances were estimated to be 2.42, .58, 22.78 and 25.78 d2, respectively, by Henderson's Method III. Heritability of gestation length was calculated to be .374 from the sire variance and .09 from the maternal grandsire variance. Direct additive genetic variance was considered to be of greater importance than maternal additive genetic variance. Correlations between the evaluations of sires for gestation length and heifer calving ease, birth weight and weaning weight were .26, .26 and .13, respectively.

Eryk - stock.adobe.com

What is a distribution

- In our case we are generally referring to a distribution function
  - This is a function (or model) that describes the probability that a system will take on value or set of values x
- For any variable X, we describe probabilities by
  - Discrete variables: probability distribution function $P(X = x)$
  - Continuous variables: probability density function $f(x)$
  - Discrete and Continuous variables: cumulative density function $F(x) = P(X \leq x)$

Properties of a continuous distribution

- For any continuous distribution
  - There is an infinite number of possible values;
  - These values may be within a fixed interval. For example, male human heights (in cm) belong to [54.6,272].

Human height

- Specific values in a continuous distribution have probability 0. For example, the likelihood of measuring a Simmental cow at exactly 450kg is zero. This is because there are potentially an infinite number of other weights that are higher or lower than 450 kg so we say that measuring exactly 450 has a very very small probability which is equivalent to zero
- The total of all the probabilities = must be 1. (Total area under the pdf)

The Normal Distribution

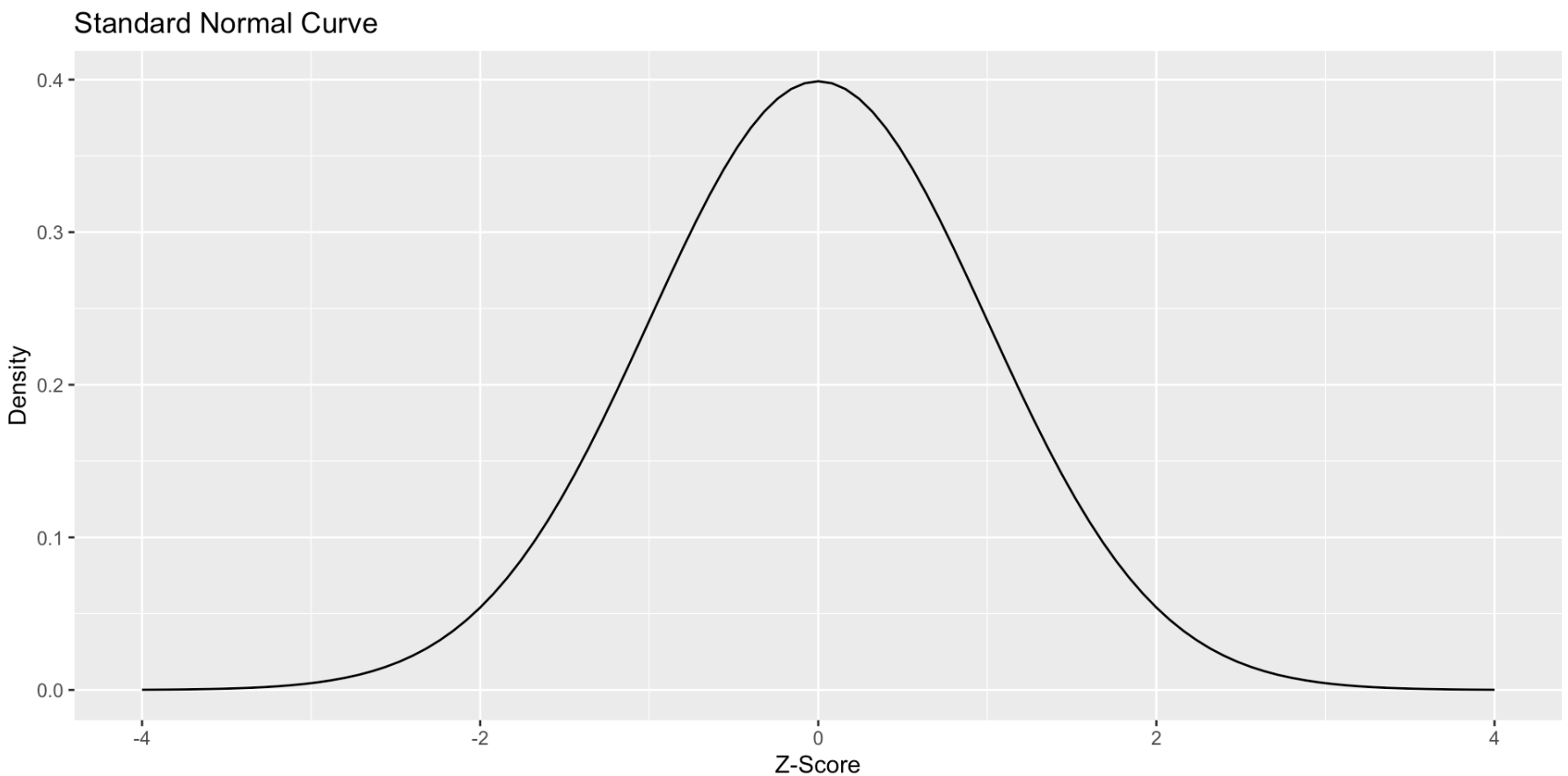
- The Normal Distribution is **super important** because it occurs everywhere! It naturally describes many natural phenomenon and is a great for modelling the sample mean.
- It is a symmetric bell-shaped variable with two parameters μ and σ^2 such that:

$$X \sim N(\mu, \sigma^2)$$

The Standard Normal Curve

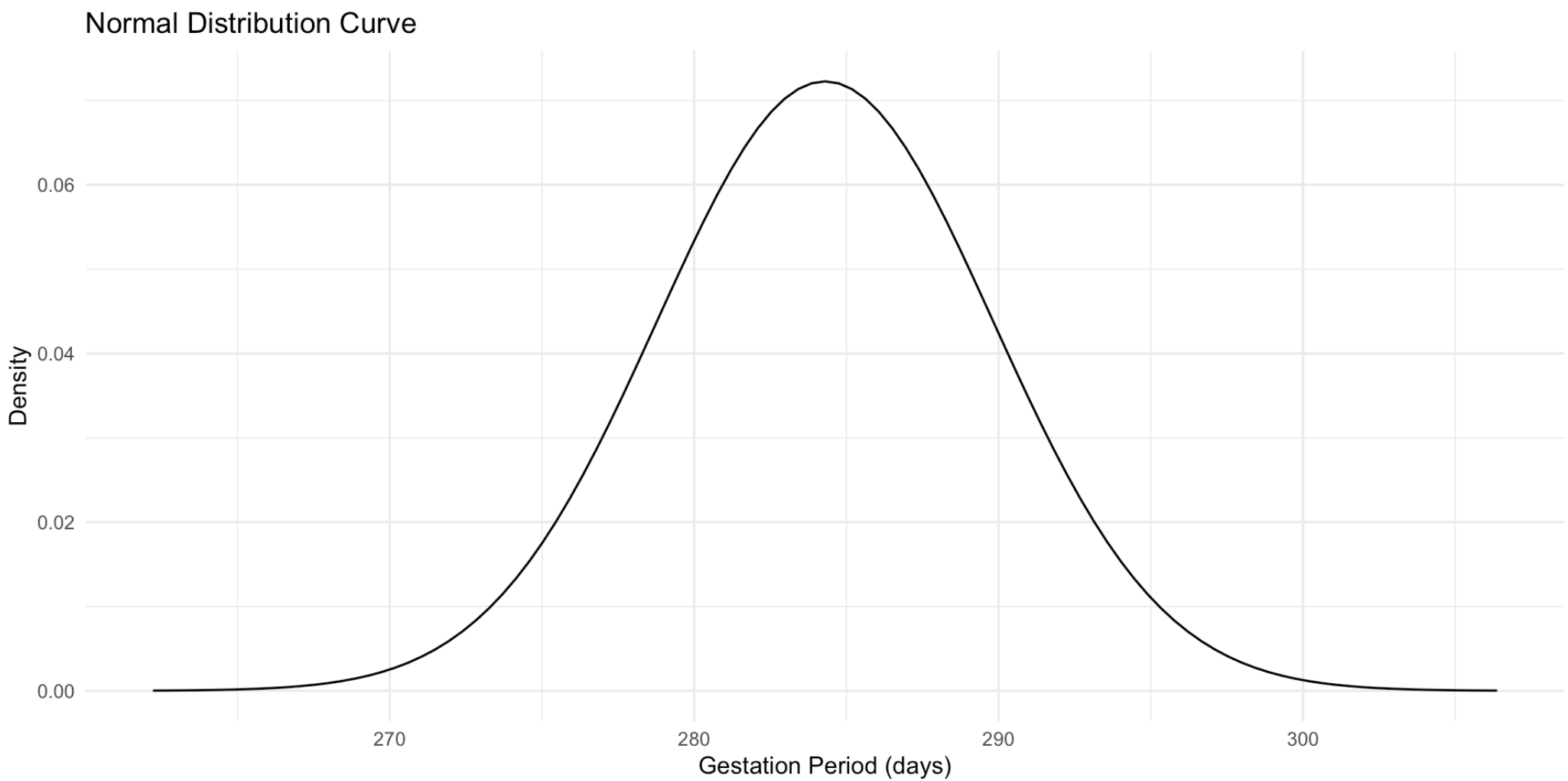
- The standard normal curve is one where the mean = 0, and variance = 1

$$X \sim N(\mu = 0, \sigma^2 = 1)$$



The General Normal Curve

- Simmental cattle gestation times…



The General Normal Distribution

If $X \sim N(\mu, \sigma^2)$

- PDF

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } x \in (-\infty, \infty)$$

- CDF

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Types of Normal Probabilitites

- There are 3 type of probabilities that we are interested in:
  - Tail probabilities (lower and upper) = Cumulative probabilities
  - Interval probabilities;
  - Inverse probabilities.

Normal distribution in R

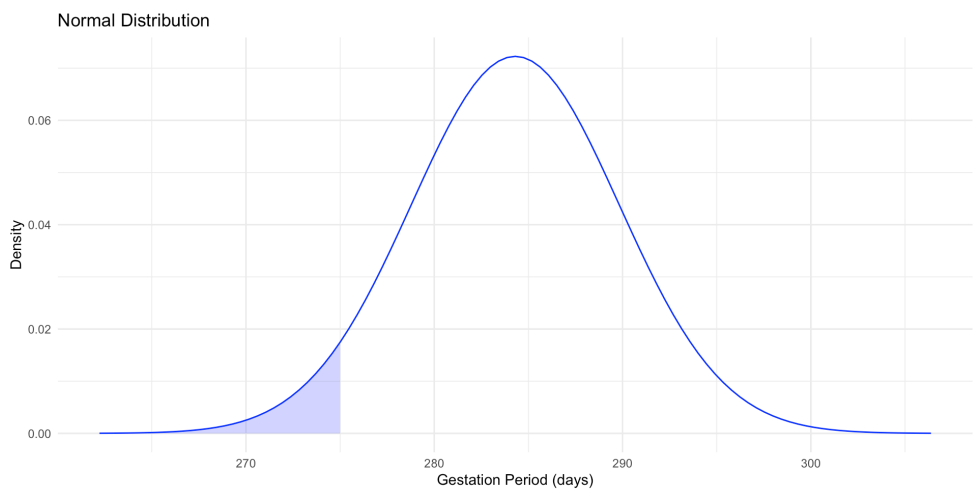
Commands for General Normal: $X \sim N(\mu, \sigma^2)$

	R
Probability density function (pdf) $f(x)$	$dnorm(x, \mu, \sigma)$
Lower Tail probabilities (cumulative, CDF) $F(x) = P(X \leq x)$	$pnorm(x, \mu, \sigma)$
Interval probabilities $P(a \leq X \leq b)$	$pnorm(b, \mu, \sigma) - pnorm(a, \mu, \sigma)$
Inverse probabilities Find q such that $P(X \leq q) = \%$	$qnorm(q, \mu, \sigma)$

Types of Normal Probabilities

Lower: $P(X \leq 275)$

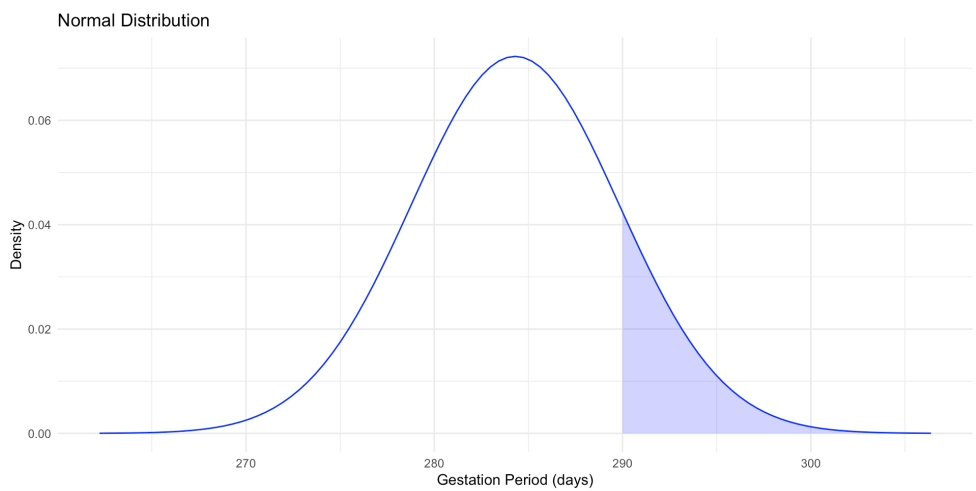
```
[1] 0.04601526
```



Types of Normal Probabilities

Upper: $P(X \geq 290)$

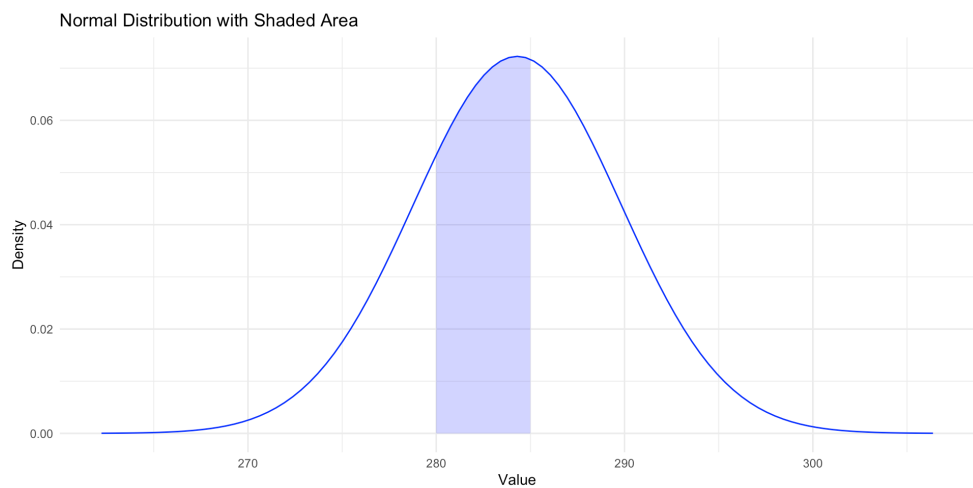
```
[1] 0.1508935
```



Types of Normal Probabilities

Interval: $P(280 \leq X \leq 285)$

```
[1] 0.3324611
```



Types of Normal Probabilities

Inverse: if we know the shaded area = 0.9 (90%), what is x?

$$P(X \leq x) = 0.9$$

```
[1] 291.3742
```

Example

- Let's return to our example for American Simmental cattle where $X \ N(284.3, 5.522)$,
- What is the probability of a gestation time less than 275 days.

So we need to calculate the lower Tail probability such that: $P(X \leq 275)$

```
[1] 0.04601526
```

- One would expect around 5% of gestation times would be less than 275 days/
- **Question for you:** Why might this be important, how can we use these results??

vxnaghiyev - stock.adobe.com

Back to the Standard Normal Curve

- Sometimes it is useful to standardise "data" as it allows us to compare samples that are drawn from populations that may have different means and standard deviations
- Luckily for us we can standardise any general normal distribution $X \sim N(\mu, \sigma^2)$ to a standard normal distribution $X \sim N(0, 1)$
- This was also useful as we could use a set of standard normal tables to calculate probabilities (before computers were readily available).

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma}, \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

Standard Normal Curve

- Example: if $X \sim N(10, 9)$ find $P(X \leq 14)$

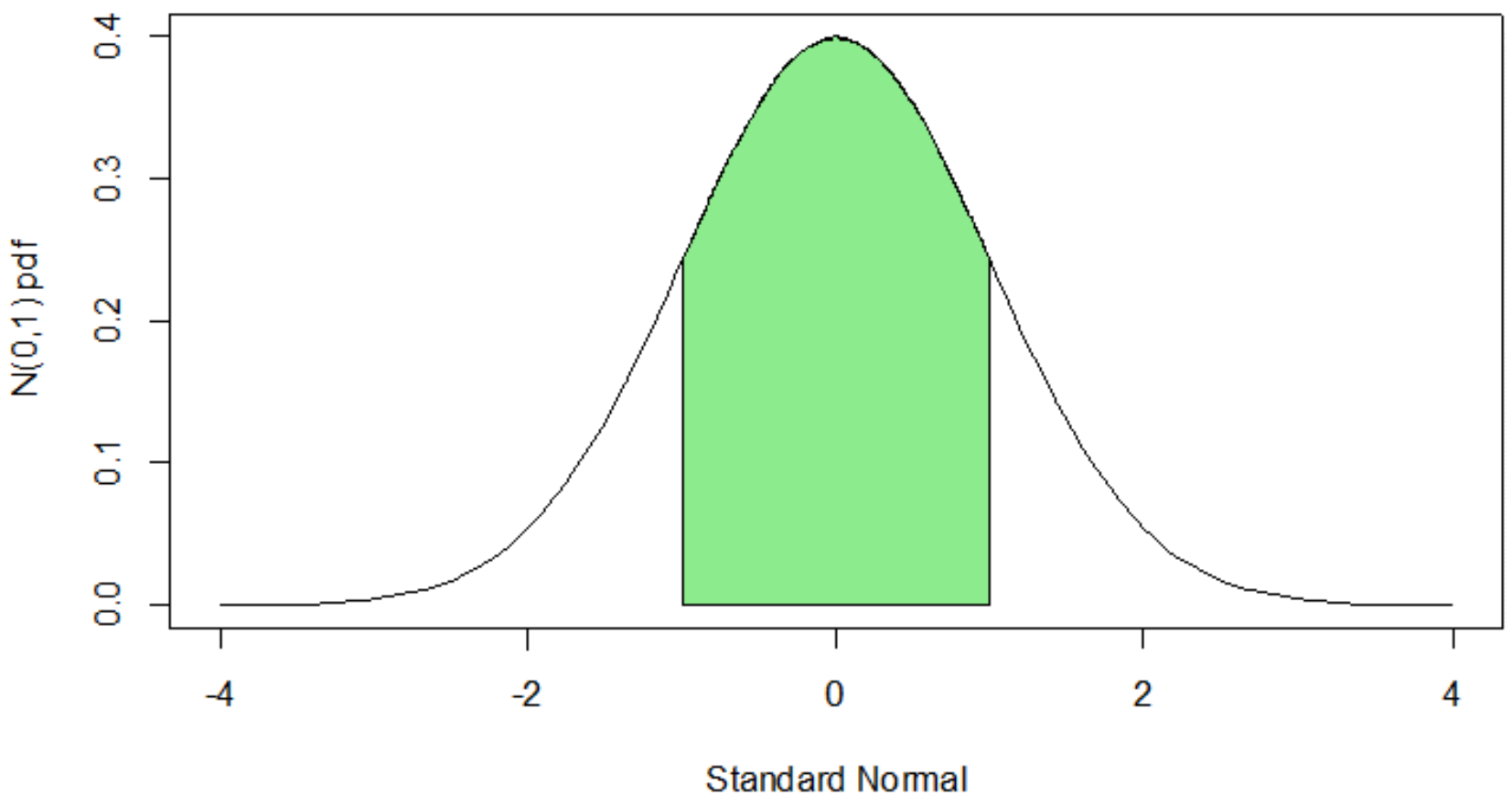
$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma}, \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

$$P(X \leq 14) = P\left(\frac{X-10}{\sqrt{9}}, \frac{14-10}{\sqrt{9}}\right) = P\left(Z \leq \frac{4}{3}\right)$$

```
[1] 0.9087888
[1] 0.9087888
[1] 0.9087888
```

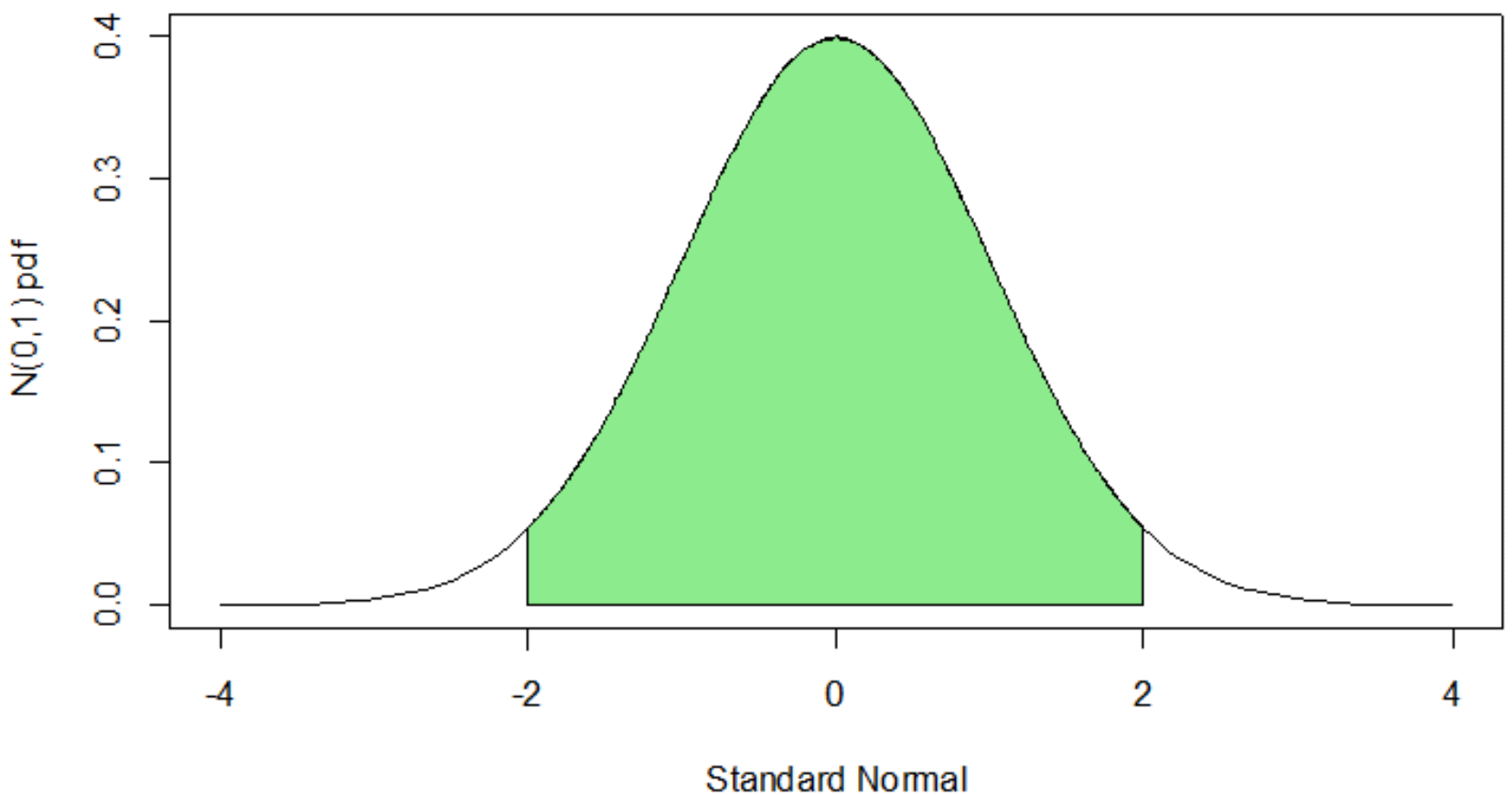
Percentiles of the Standard Normal Curve

- 1 Standard deviation from the mean = 68% of the data



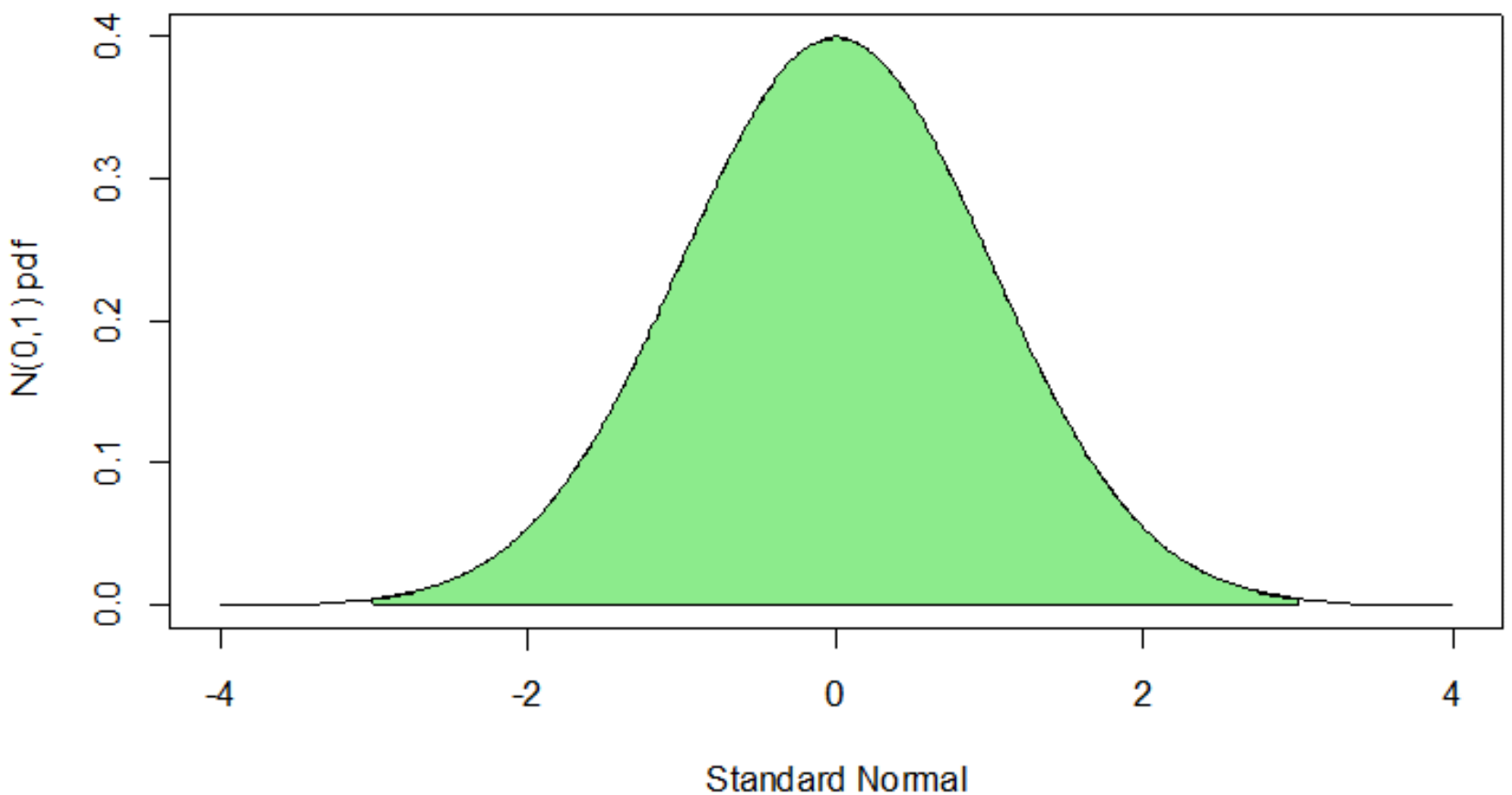
Percentiles of the Standard Normal Curve

- 2 Standard deviations from the mean = 95% of the data



Percentiles of the Standard Normal Curve

- 3 Standard deviations from the mean = 99.7% of the data



Not so Normal Distributions

Not so Normal Distributions

Sampling distributions

- Rye grass root growth (in mg dry weight) follows the distribution $X \sim N(300, 502)$.
- 1. One measurement is taken: how likely is it that the dry weight exceeds 320 mg?
- 2. 10 measurements are taken: how likely is it that the sample mean exceeds 320 mg?

SERHII BLIK - stock.adobe.com

Sampling distributions

- Here, we are dealing with 2 distributions:
  1. Measurement: $X \sim N(300, 50^2)$
  2. Sample Mean of 10 measurements: $\overline{X} = \frac{1}{10}\Sigma_{i=1}^{10} X_i \sim ...$

How does the sampling distribution occur?

- http://onlinestatbook.com/stat_sim/sampling_dist/
- We have a population X
  - We take a sample of size n and we calculate the mean $\overline{x}_1$
  - We take another sample of size n and we calculate the mean $\overline{x}_2$
  - We take another sample of size n and we calculate the mean $\overline{x}_3$ … If we sample all possibilities, then the sampling distribution of $\overline{X} = \frac{1}{10}\Sigma_{i=1}^{10} X_i$ is the distribution of $\{\overline{x}_1, \overline{x}_2, \overline{x}_3, ...\}$

Distribution for a sample mean

- if $X \sim N(\mu, \sigma^2)$
- then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Note that we call
  - σ the standard deviation such that $sd(X) = \sigma$, and
  - $\sigma/\sqrt{n}$ the standard error such that $sd(\overline{X}) = \sigma/\sqrt{n}$
- The standard error is important for making inference on a sample populations i.e. how close your sample mean $\overline{x}$ is to the population mean μ

Example

- Rye grass root growth (in mg dry weight) follows the distribution $X \sim N(300, 50^2)$.
    1. One measurement is taken: how likely is it that the dry weight exceeds 320 mg?
    2. 10 measurements are taken: how likely is it that the sample mean exceeds 320 mg?

Example

- 1. X = Rye grass root growth $\sim N(300, 50^2)$

$P(X > 320) = P\left(\frac{X-\mu}{\sigma}, \frac{x-\mu}{\sigma}\right) = P\left(\frac{X-300}{50} > \frac{320-300}{50}\right) = P(Z > 0.4) = 1 - P(Z < 0.4) \approx 1 - 0.66 = 0.34$

```
[1] 0.3445783
```

Example

- 1. $\overline{X}$ = Rye grass root growth $\sim N(300, \frac{50^2}{10})$

$$P(\overline{X} > 320) = P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}, \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\frac{\overline{X}-300}{\frac{50}{\sqrt{10}}} > \frac{320-300}{\frac{50}{\sqrt{10}}}\right) = P(Z > 1.26) = 1 - P(Z < 1.26)$$

$$\approx 1 - 0.90 = 0.10$$

```
[1] 0.1038347
```

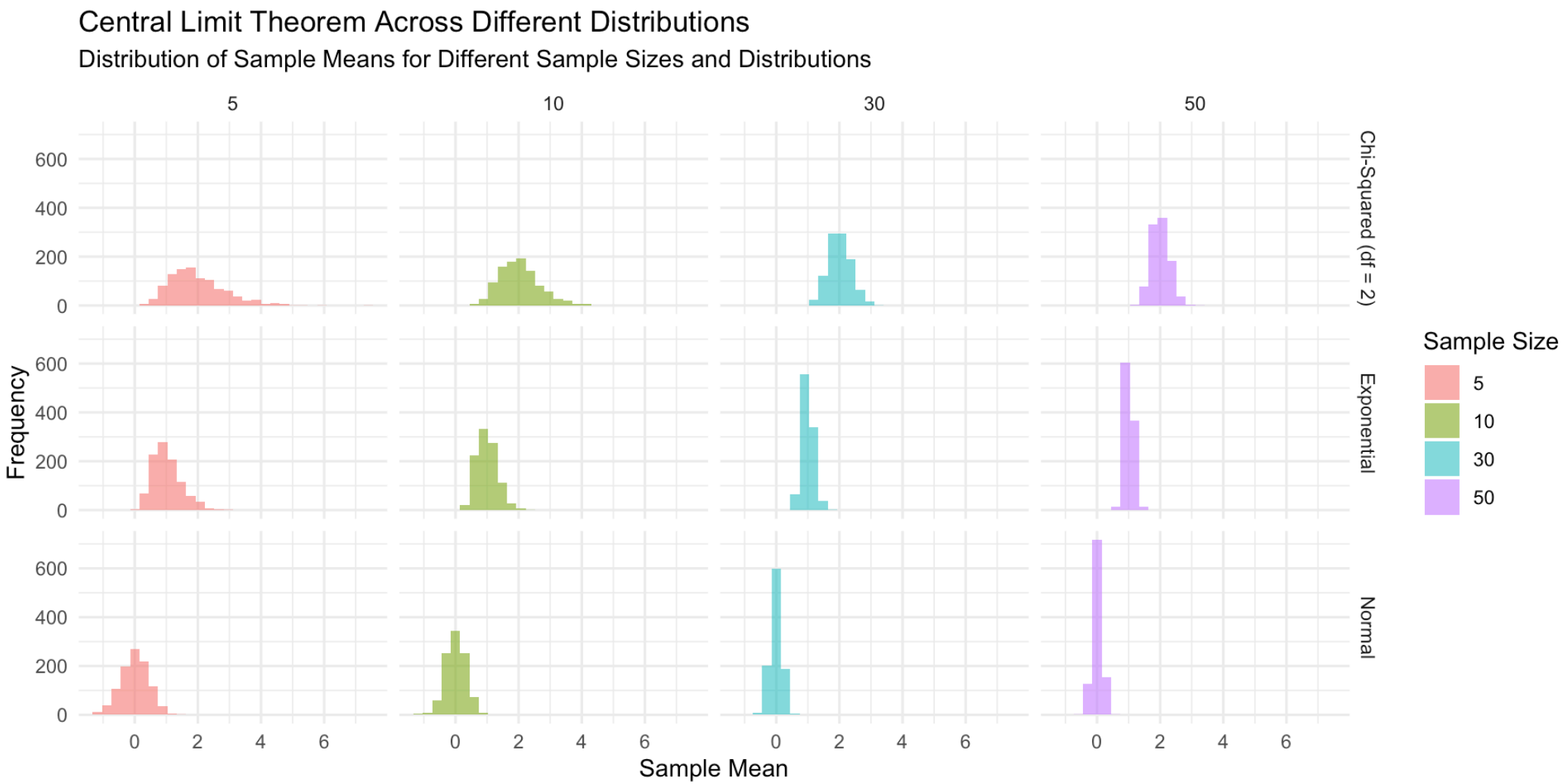
Central limit theorem

- what about if we sample from any old population such that $X \sim ??(\mu, \sigma^2)$?
- The it follows that $X \approx N(\mu, \sigma^2)$) and this is as a result of the Central Limit Theorem
- The CLT is the most important result in this course, and in much of statistical theory.
- The CLT requires few assumptions:
  - We must have a `big enough' sample size n;
  - We must have finite variance $\sigma^2 < \infty$ What is a `big enough' sample size? Some textbooks give a rule of thumb (eg $n > 25$ or $n > 30$), but it all depends on the type of distribution. If X is reasonably symmetric, then n could be small; if X is highly asymmetric, then n could be larger.

Central limit theorem

- We can demonstrate this in R (you will do it in the practical)

Central limit theorem



Thanks!

title: “ENVX1002-2024-Lecture-Topic04” format: revealjs editor: visual