Lecture 10a: Confidence intervals (CIs)

Question: Write pseudocode for the bootstrap method for estimating standard errors.

Worksheets + scratch paper + snacks are in front!

Josh Grossman

shorturl.at/rt5m6

📓 Logistical notes

HW2 now due Monday at midnight.

[I want to make sure we have sufficiently covered CIs]

📓 Logistical notes

HW2 now due Monday at midnight.
[I want to make sure we have sufficiently covered CIs]

Lab 4 (Parts A + B) w/ Lecture 9 worksheet due next Friday at midnight.

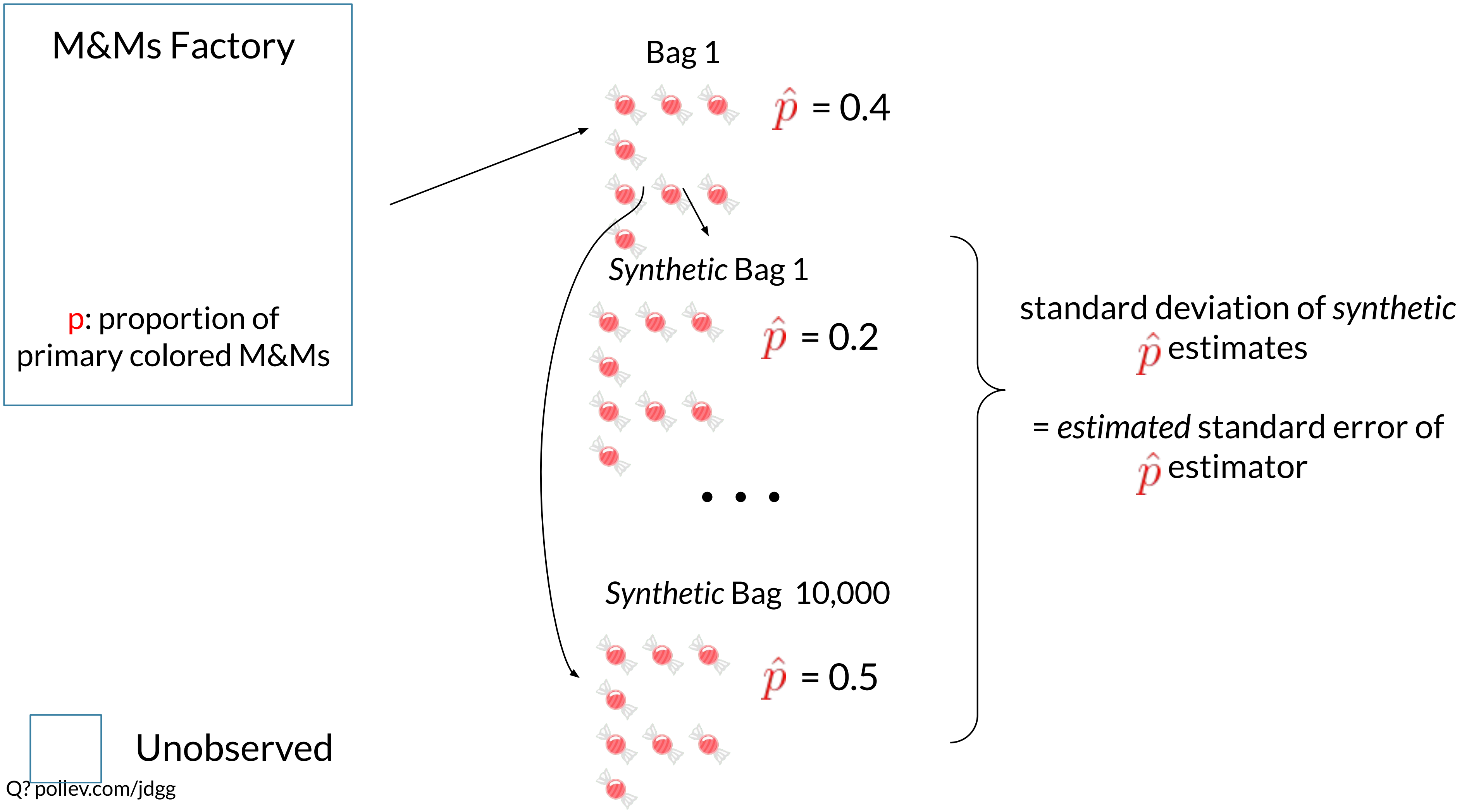
📓 Logistical notes

HW2 now due Monday at midnight.
[I want to make sure we have sufficiently covered CIs]

Lab 4 (Parts A + B) w/ Lecture 9 worksheet due next Friday at midnight.

Midterm 1 is two weeks from today.
[Practice problems posted]

Midterm logistics post coming soon.



M&Ms Factory
p: proportion of primary colored M&Ms
Unobserved
Bag 1
$\hat{p}$ = 0.4
Synthetic Bag 1
$\hat{p}$ = 0.2
Synthetic Bag 10,000
$\hat{p}$ = 0.5
standard deviation of *synthetic* $\hat{p}$ estimates
= *estimated* standard error of $\hat{p}$ estimator
Q? pollev.com/jdgg

The bootstrap

0. Start with a single random sample of size n. Calculate a point estimate using this sample.

The bootstrap

0. Start with a single random sample of size n. Calculate a point estimate using this sample.

1. Resample with replacement from the original sample to produce B synthetic samples, each of size n.

🥾 The bootstrap

0. Start with a single random sample of size n. Calculate a point estimate using this sample.

1. Resample with replacement from the original sample to produce B synthetic samples, each of size n.

2. Calculate a synthetic point estimate for each of the B synthetic samples.

🥾 The bootstrap

0. Start with a single random sample of size n. Calculate a point estimate using this sample.

1. Resample with replacement from the original sample to produce B synthetic samples, each of size n.

2. Calculate a synthetic point estimate for each of the B synthetic samples.

3. The SD of the B synthetic point estimates is the bootstrap-estimated SE.

Presidential polling from April 2024

From Reuters/Ipsos

41% of registered voters in the April 5 – April 9 poll said they would vote for Biden, compared with 37% who picked Trump.

Presidential polling from April 2024

From Reuters/Ipsos

41% of registered voters in the April 5 – April 9 poll said they would vote for Biden, compared with 37% who picked Trump.

Responses were from **833 registered voters** surveyed online.

Presidential polling from April 2024

From Reuters/Ipsos

41% of registered voters in the April 5 – April 9 poll said they would vote for Biden, compared with 37% who picked Trump.

Responses were from **833 registered voters** surveyed online.

The poll had a **margin of error of 4%** with **95% confidence**.

Presidential polling from April 2024

From Reuters/Ipsos

41% of registered voters in the April 5 – April 9 poll said they would vote for Biden, compared with 37% who picked Trump.

Responses were from **833 registered voters** surveyed online.

The poll had a **margin of error of 4%** with **95% confidence**.

What's your interpretation of the 4% and 95%? How might a non-technical consumer of news interpret 4% and 95%?
[Discuss with neighbors]

🥅 Confidence intervals (CIs)

A 95% confidence interval (CI) for a parameter θ is an interval $C_n = (a,b)$ such that:

$$\Pr(\theta \in C_n) \geq 0.95$$

🥅 Confidence intervals (CIs)

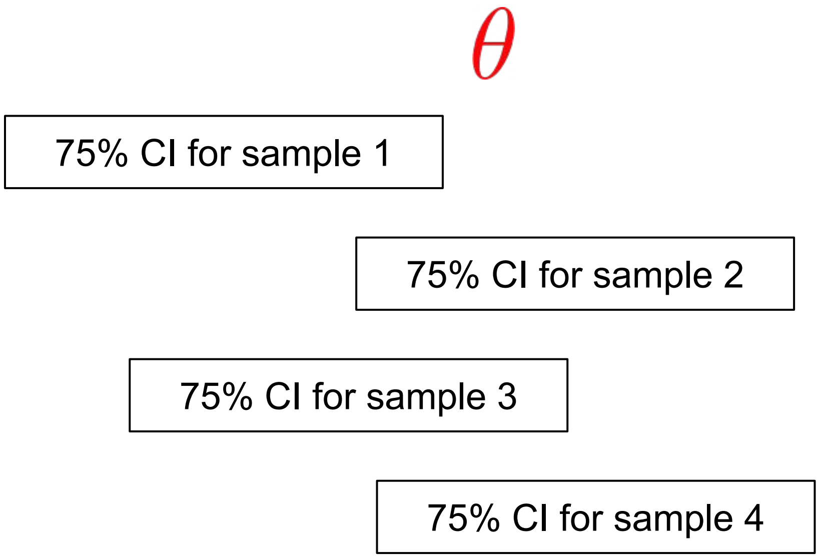
A 95% confidence interval (CI) for a parameter θ is an interval C_n = (a,b) such that:

$$\Pr(\theta \in C_n) \geq 0.95$$

θ is fixed and C_n is random

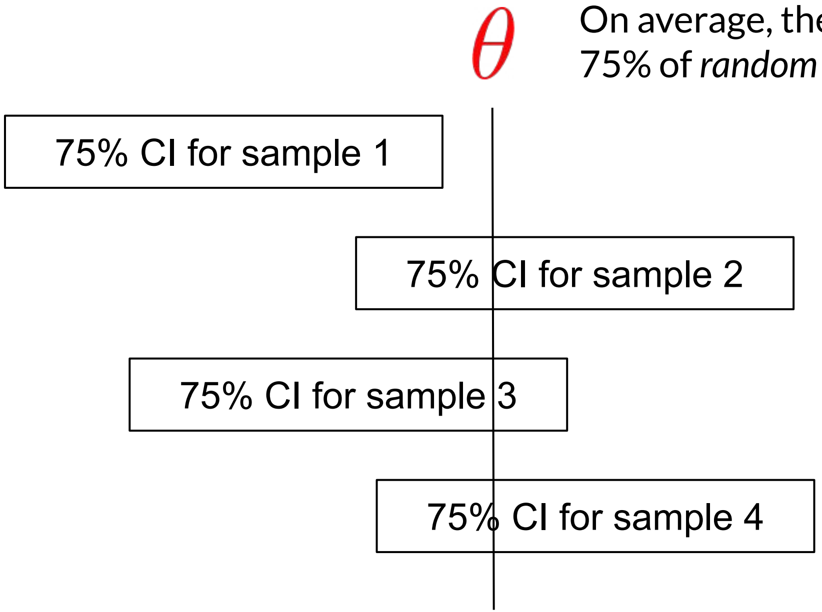
"We're not shooting soccer balls at stationary goals. We're trying to place goal posts around stationary soccer balls."

🥅 Confidence intervals (CIs)



🥅 Confidence intervals (CIs)

On average, the true *fixed* θ falls in 75% of *random* CIs



🥅 Confidence intervals (CIs)

Two interpretations

1. If you repeat the **same** experiment many times, the 95% CIs will "capture" the true parameter at least 95% of the time.

But, we only get to observe one universe!

🥅 Confidence intervals (CIs)

Two interpretations

1. If you repeat the **same** experiment many times, the 95% CIs will "capture" the true parameter at least 95% of the time.

But, we only get to observe one universe!

2. If you construct many 95% CIs for various, **unrelated** parameters, at least 95% of CIs will contain their parameter.

🥅 Confidence intervals (CIs)

Examples

49% ± 3% think U.S. should lift Cuba embargo.

38% ± 3% think U.S. should build more nuclear power plants.

16% ± 4% think St. Louis Cardinals will win the World Series.

🥅 Confidence intervals (CIs)

Examples

49% ± **3%** think U.S. should lift Cuba embargo.

38% ± **3%** think U.S. should build more nuclear power plants.

16% ± **4%** think St. Louis Cardinals will win the World Series.

Margins of error

Point estimates
[Best guesses]

🥅 Confidence intervals (CIs)

A 95% confidence interval (CI) for a parameter θ is an interval C_n = (a,b) such that:

$$\Pr(\theta \in C_n) \geq 0.95$$

🏗️ Constructing intervals

A $1-\alpha$ confidence interval (CI) for a parameter θ is an interval $C_n = (a,b)$ such that:

$$\Pr(\theta \in C_n) \geq 1 - \alpha$$

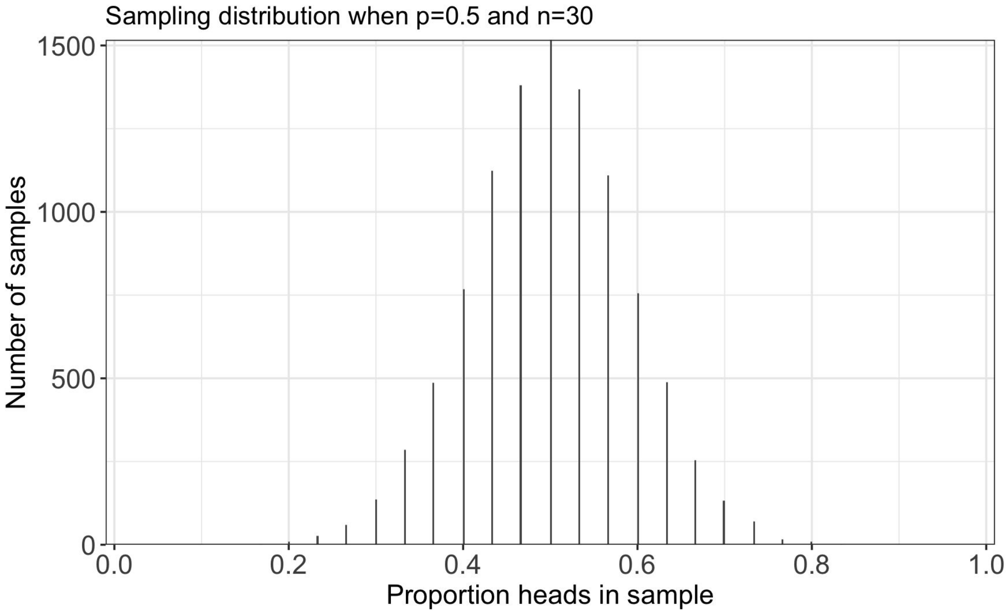
🏗️ Constructing intervals

Normal approximations

Suppose, by the CLT,

$$\hat{\theta}_n \approx N(\theta, \mathrm{se}^2)$$

Flip 30 coins (p=0.5) 10,000 times, and plot the 10,000 resulting means.



Across parallel universes of random samples, what estimates might we have observed?

This is the approximate **sampling distribution** of our $\hat{p}$ estimator when p=0.5 and n=30.

The true sampling distribution has infinite replications → We treat 10,000 as ∞ !

🏗️ Constructing intervals

Normal approximations

Suppose, by the CLT,

$$\hat{\theta}_n \approx N(\theta, \mathrm{se}^2)$$

Let

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\mathrm{se},\ \hat{\theta}_n + z_{\alpha/2}\mathrm{se})$$

🏗️ Constructing intervals

Normal approximations

Suppose, by the CLT,

$$\hat{\theta}_n \approx N(\theta, \boxed{\text{se}}^2)$$

Let

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\boxed{\text{se}},\ \hat{\theta}_n + z_{\alpha/2}\boxed{\text{se}})$$

If possible, estimate the analytic standard error,
OTHERWISE,
bootstrap the standard error

🏗️ Constructing intervals

Normal approximations

Suppose, by the CLT,

$$\hat{\theta}_n \approx N(\theta, \mathrm{se}^2)$$

Let

?   ?

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\mathrm{se},\ \hat{\theta}_n + z_{\alpha/2}\mathrm{se})$$

🏗️ Constructing intervals

[See skipped Lecture 10b slides for short proof! Not required for 131A.]

Suppose, by the CLT,

$$\hat{\theta}_n \approx N(\theta, \text{se}^2)$$

Let

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\text{se},\ \hat{\theta}_n + z_{\alpha/2}\text{se})$$

(? ?)

Then

$$\Pr(\theta \in C_n) \approx 1 - \alpha$$

General Q&A

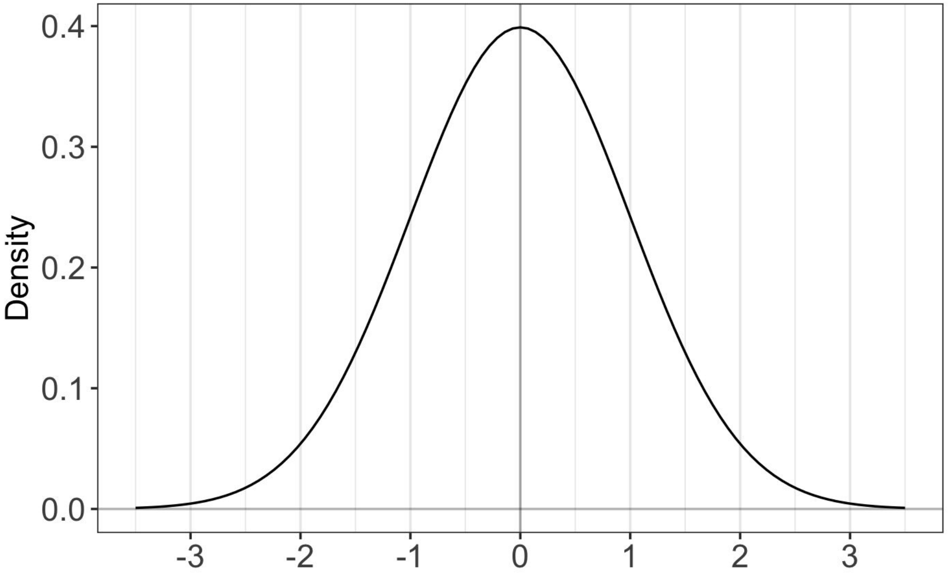
Nobody has responded yet.

Hang tight! Responses are coming in.

The normal distribution

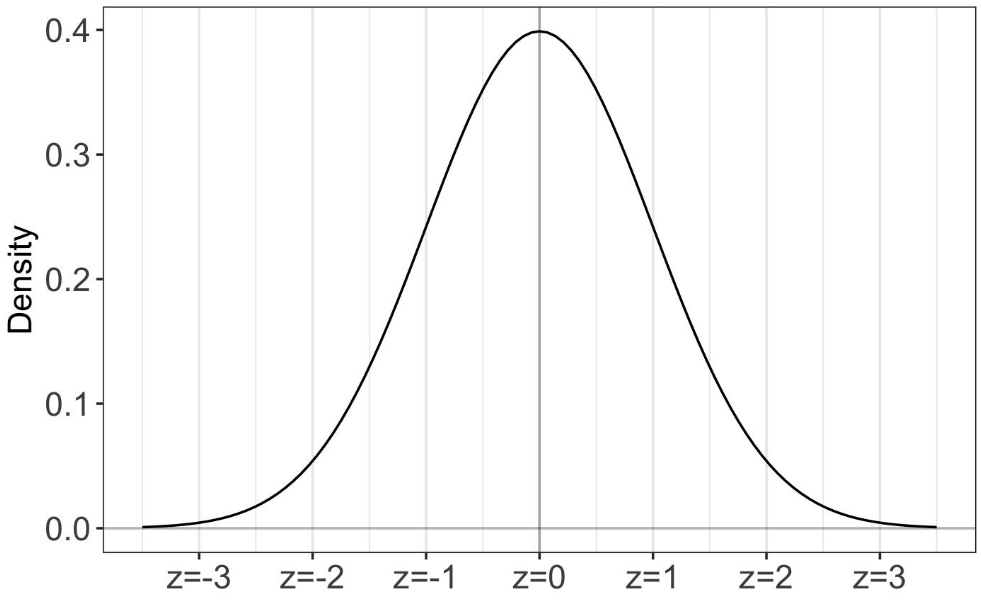
🔔 The standard normal distribution

Mean=0, SD=1, also known as the z-distribution



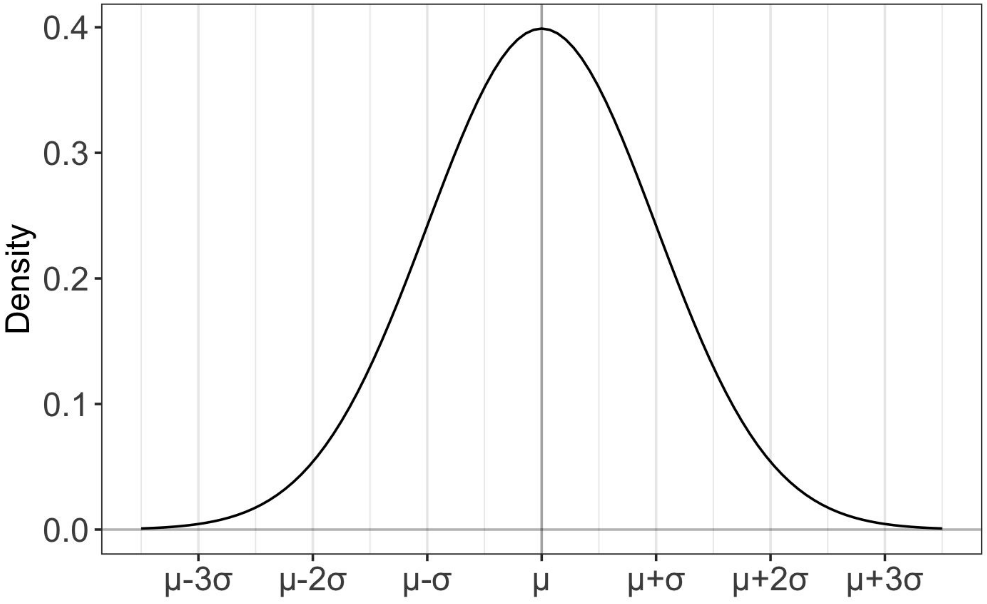
🦓 Z is the number of SDs from the mean

Also known as z-score

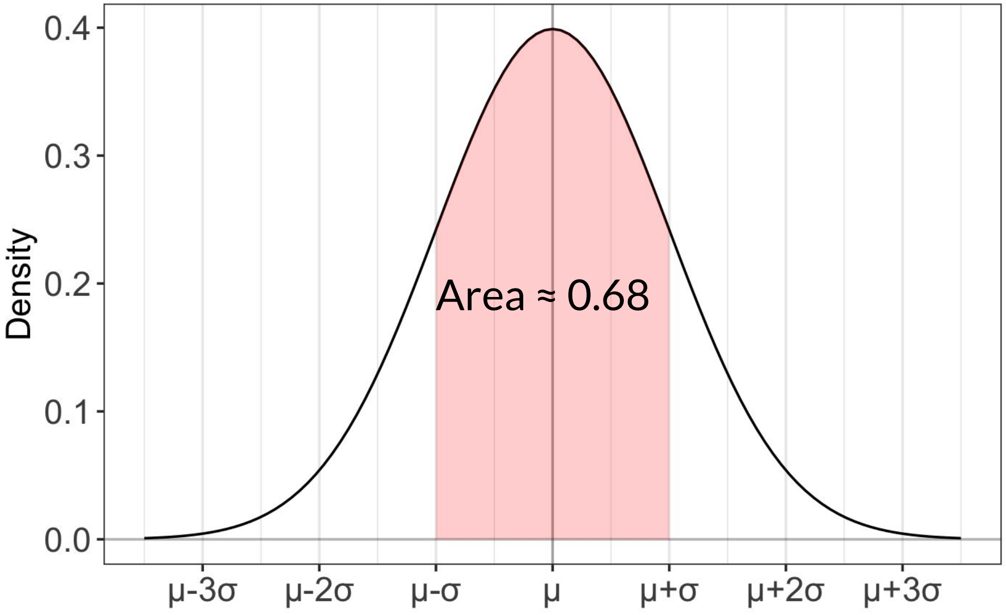


🔔 An arbitrary normal distribution

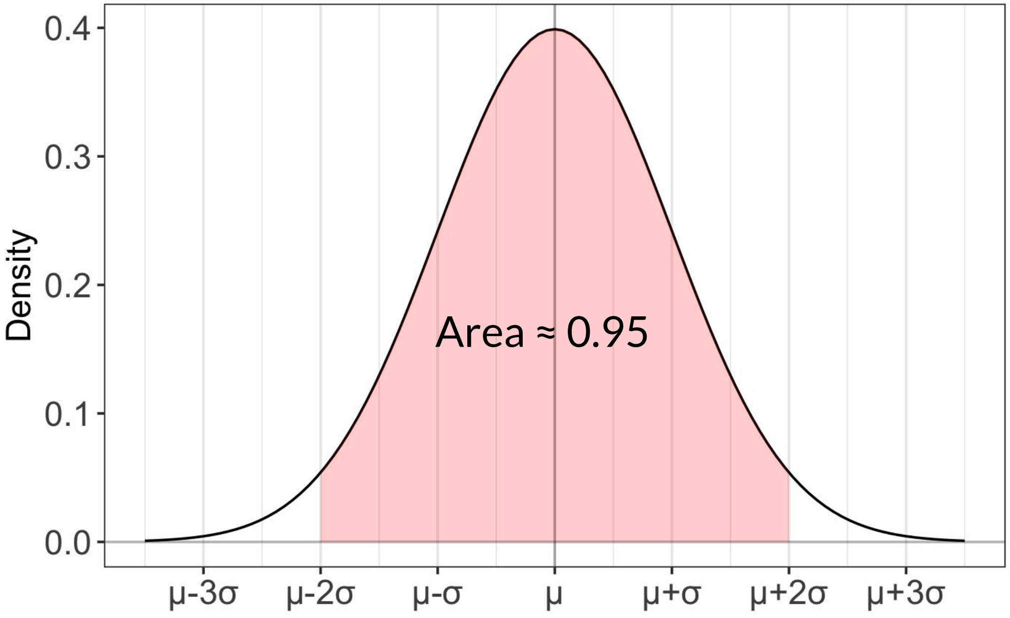
With mean of μ and SD of σ



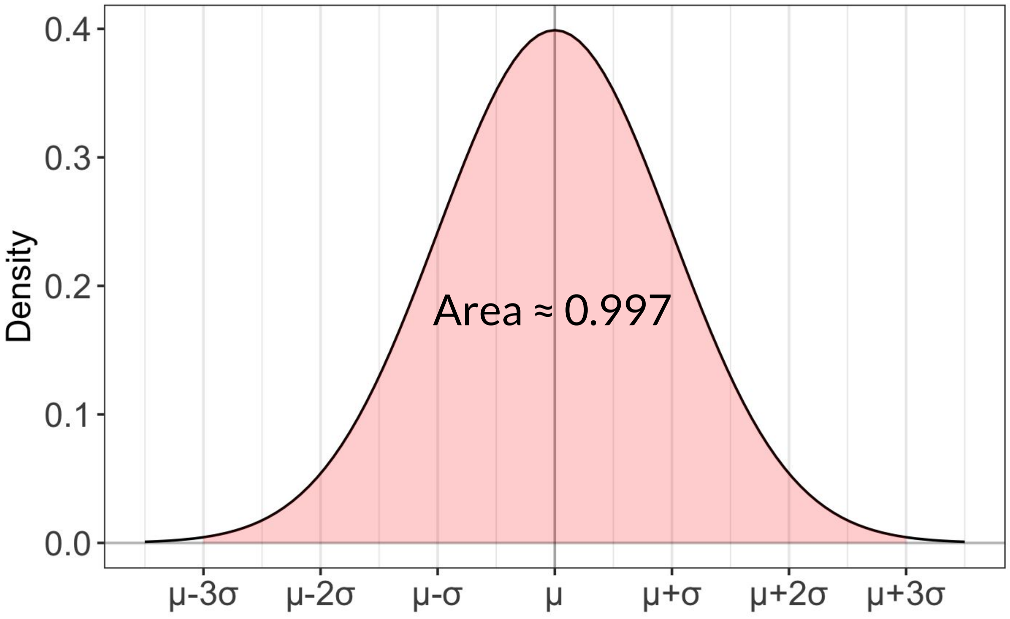
🔔 68-95-99.7 rule



🔔 68-95-99.7 rule

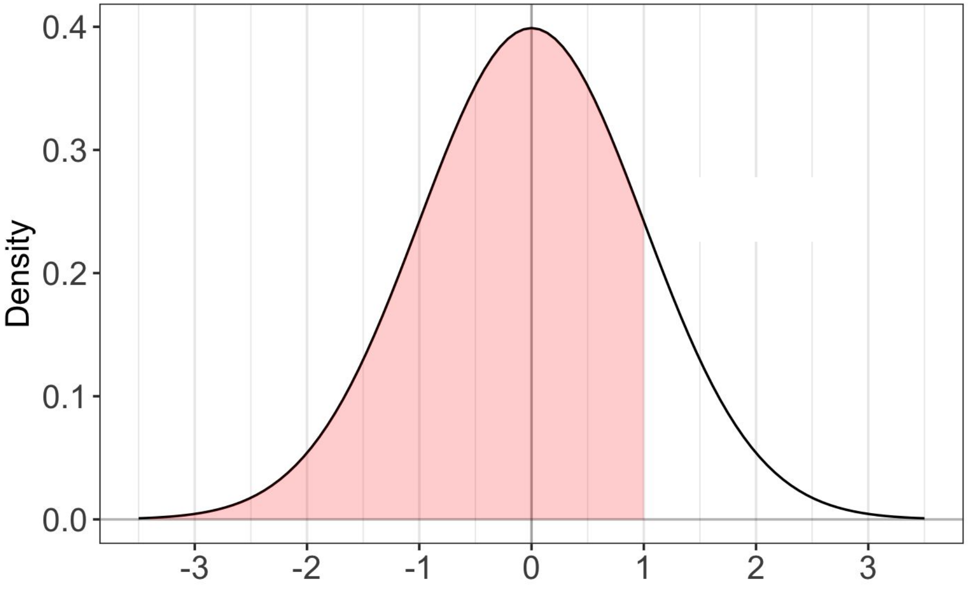


🔔 68-95-99.7 rule



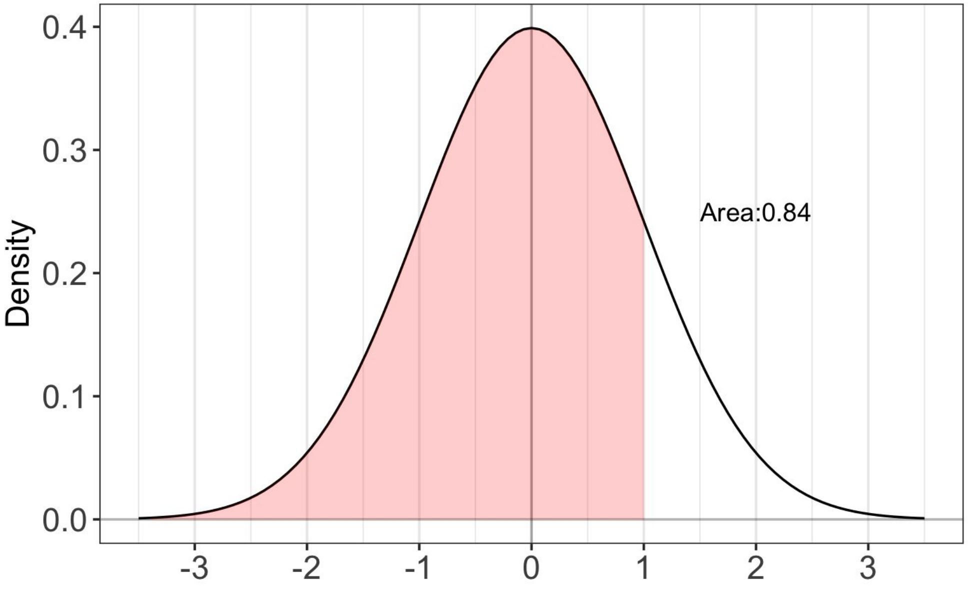
🐟 Left-tailed area

What is Pr(Z < 1)? [Worksheet]

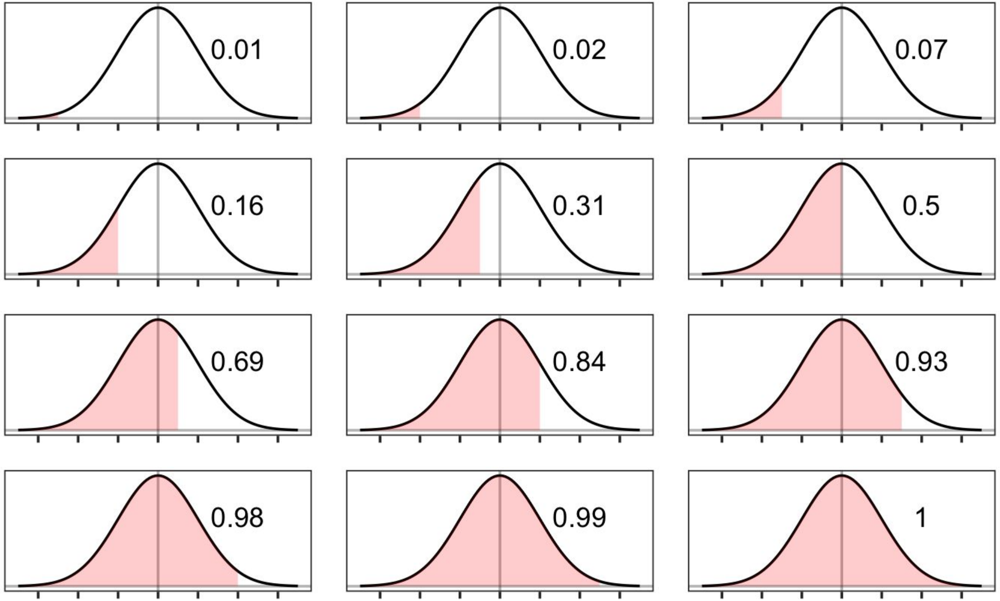


🐟 Left-tailed area

What is Pr(Z < 1)? [Worksheet]

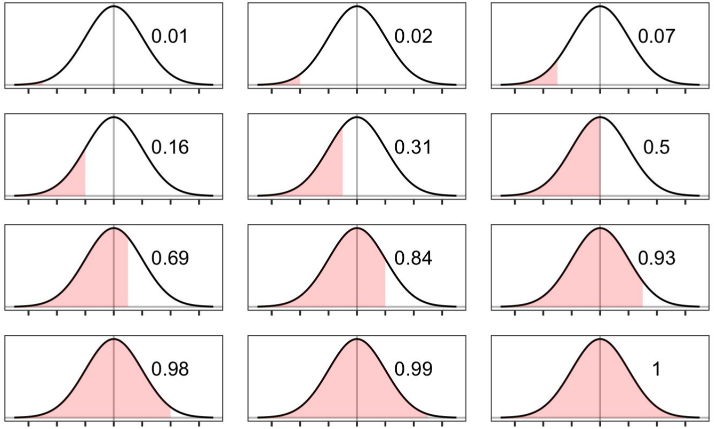


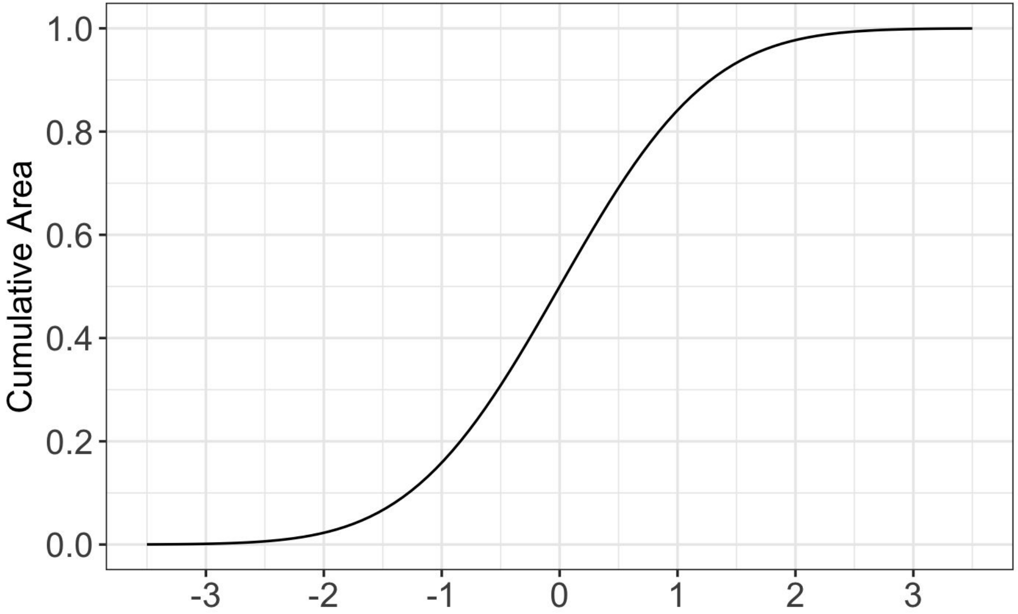
🐟 Left-tailed area



The cumulative distribution function (CDF)

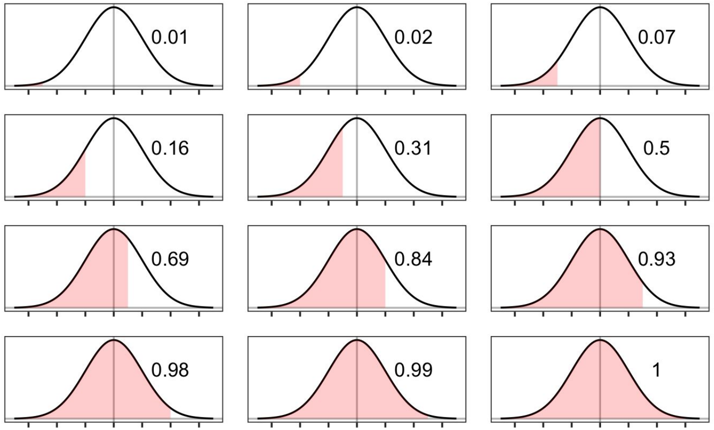
For the normal distribution

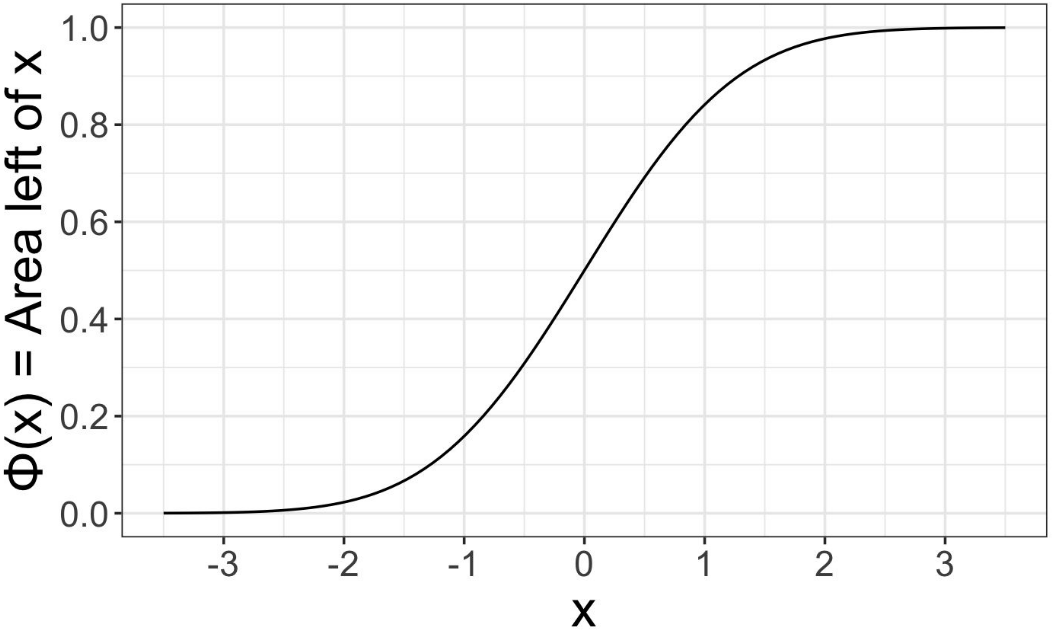




The cumulative distribution function (CDF)

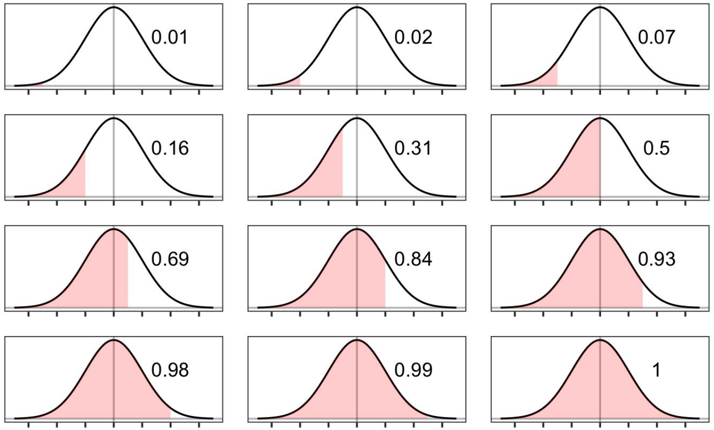
For the normal distribution





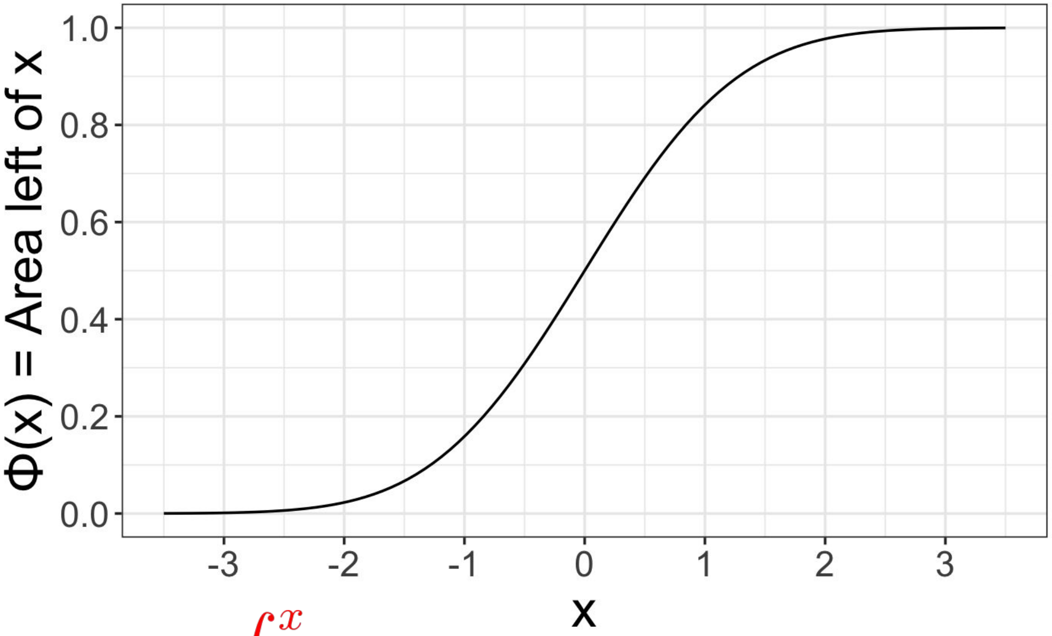
The cumulative distribution function (CDF)

For the normal distribution



Normal PDF

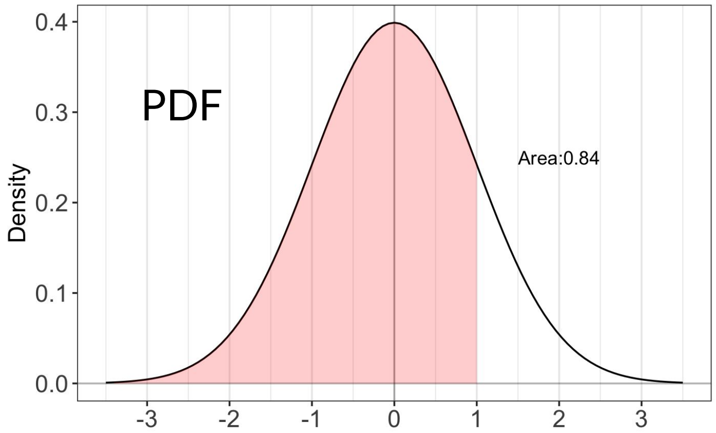
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

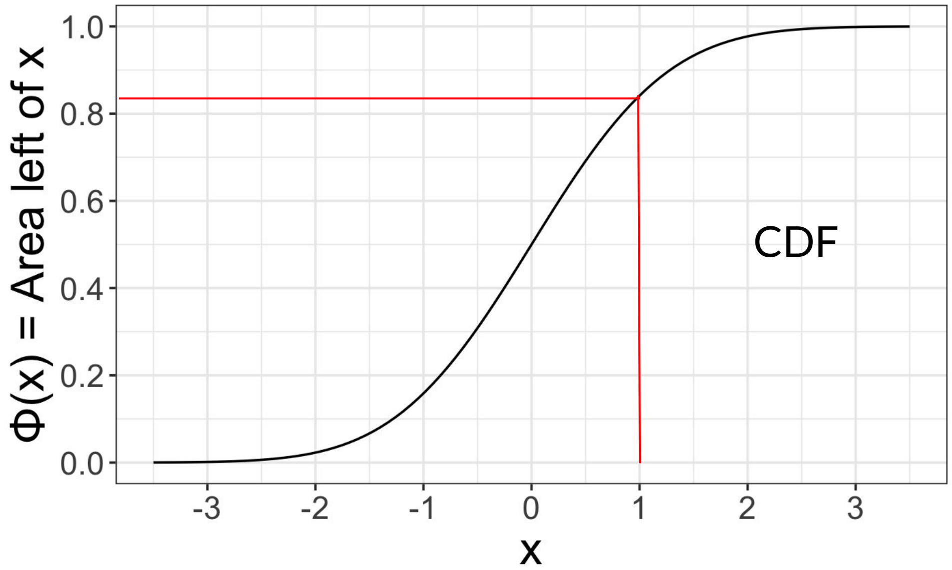


$$\Phi(x) = \int_{-\infty}^{x} f(x)dx$$

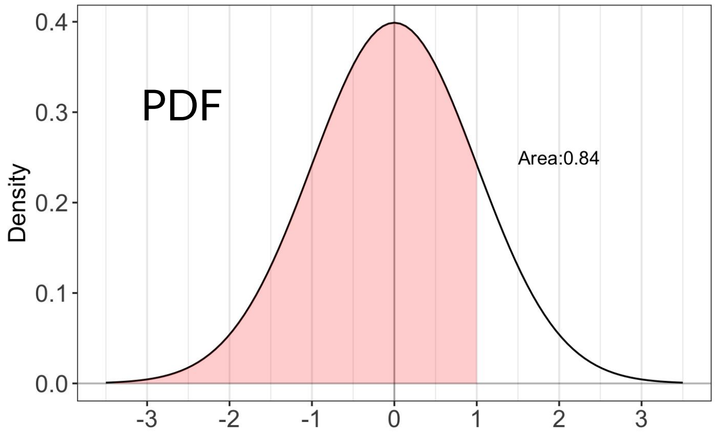
Normal CDF

Q? pollev.com/jdgg

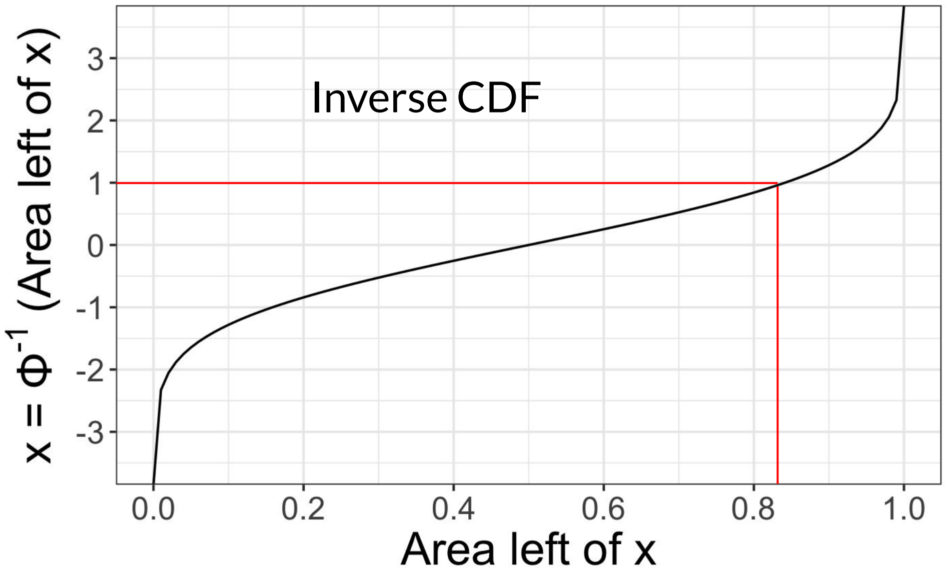

PDF
Area:0.84
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

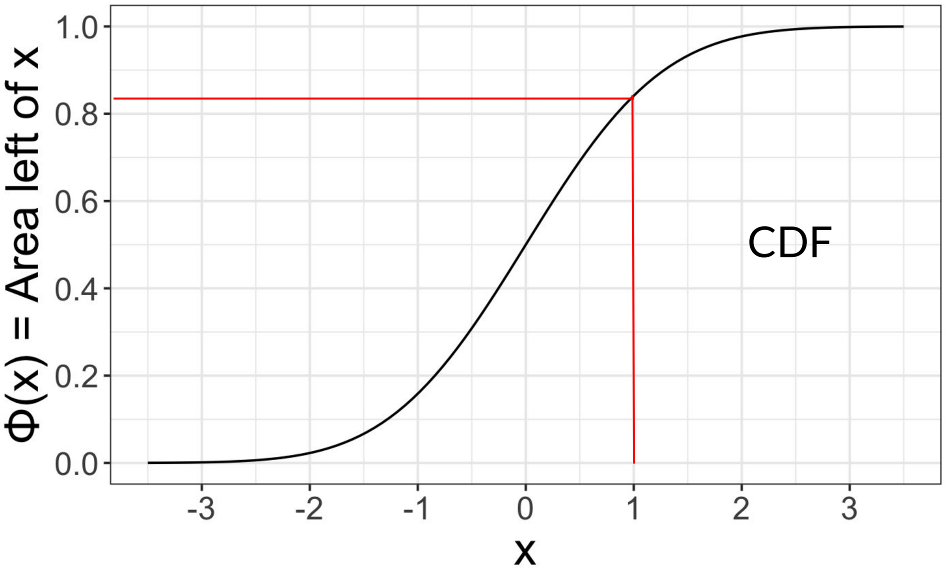


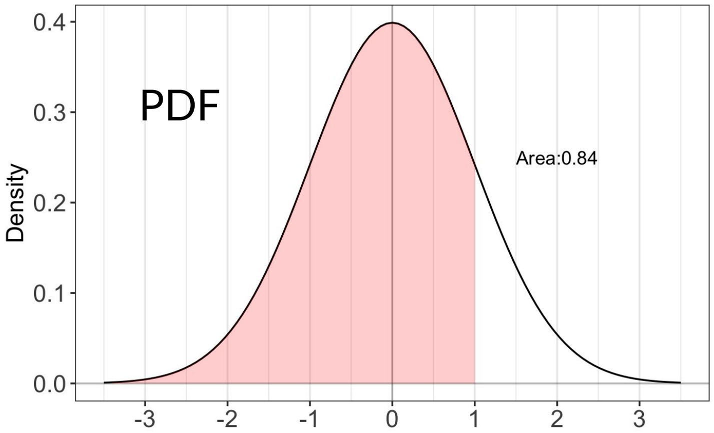
Φ(x) = Area left of x
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
x
CDF



Density
0.4
0.3
0.2
0.1
0.0
PDF
Area:0.84
-3
-2
-1
0
1
2
3


Inverse CDF
x = Φ-1 (Area left of x)
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.0
Area left of x


CDF
Φ(x) = Area left of x
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
x


PDF
Area:0.84
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

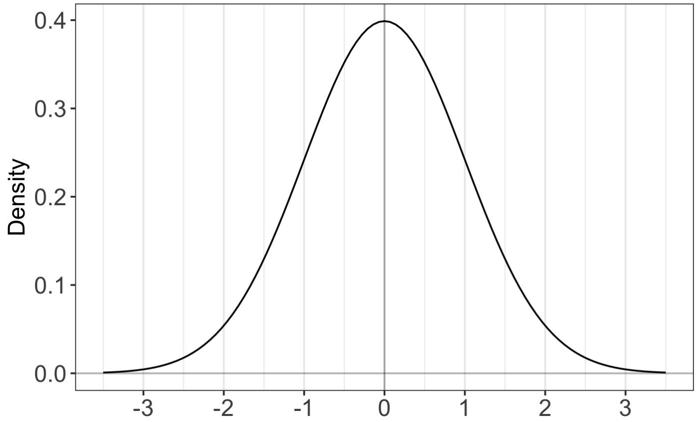
🏗️ Constructing intervals

Normal approximations

Let Φ be the CDF of a standard normal N(0,1)

What is $\Phi^{-1}(0.025)$?

[📝 Worksheet]



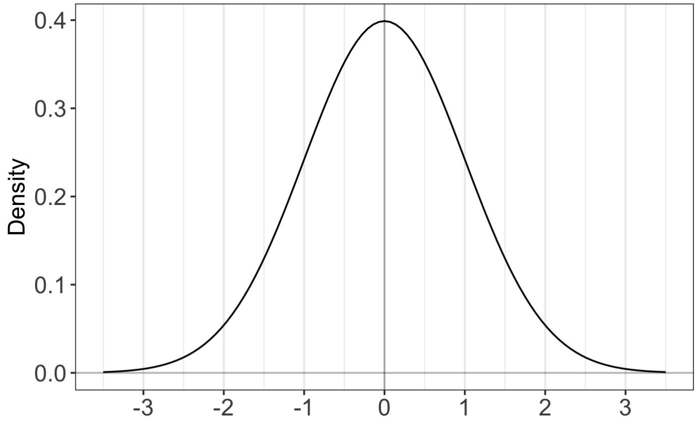
🏗️ Constructing intervals

Normal approximations

Let Φ be the CDF of a standard normal $N(0,1)$

What is $\Phi^{-1}(0.025)$?

"What x-axis value corresponds to a left-tailed area of 0.025?"



Q? pollev.com/jdgg

🏗️ Constructing intervals

Normal approximations

Let Φ be the CDF of a standard normal N(0,1)

$$\Phi^{-1}(0.025) = -2$$

$$\Phi^{-1}(0.975) = +2$$

[Technically, -1.96 and 1.96]