2.1 — Random Variables & Distributions

ECON 480 • Econometrics • Fall 2021

Ryan Safner

Assistant Professor of Economics

safner@hood.edu

[ryansafner/metricsF21](https://github.com/ryansafner/metricsF21)

metricsF21.classes.ryansafner.com

Random Variables

Experiments

- An **experiment** is any procedure that can (in principle) be repeated infinitely and has a well-defined set of outcomes

Example: flip a coin 10 times

Random Variables

- A **random variable (RV)** takes on values that are unknown in advance, but determined by an experiment
- A numerical summary of a random outcome

Example: the number of heads from 10 coin flips

Random Variables: Notation

- Random variable X takes on individual values (x_i) from a set of possible values
- Often capital letters to denote RV's
  - lowercase letters for individual values

Example: Let X be the number of Heads from 10 coin flips. $x_i \in \{0, 1, 2, \ldots, 10\}$

Discrete Random Variables

Discrete Random Variables

- A **discrete random variable**: takes on a finite/countable set of possible values

Example: Let X be the number of times your computer crashes this semester[1], $x_i \in \{0, 1, 2, 3, 4\}$

[1] Please, back up your files!

Discrete Random Variables: Probability Distribution

- **Probability distribution** of a R.V. fully lists all the possible values of X and their associated probabilities

Example:

x_i	$P(X = x_i)$
0	0.80
1	0.10
2	0.06
3	0.03
4	0.01

Discrete Random Variables: pdf

Probability distribution function (pdf) summarizes the possible outcomes of X and their probabilities

- Notation: f_X is the pdf of X:

$$f_X = p_i, \qquad i = 1, 2, \ldots, k$$

- For any real number x_i, $f(x_i)$ is the probablity that $X = x_i$
- What is $f(0)$?
- What is $f(3)$?

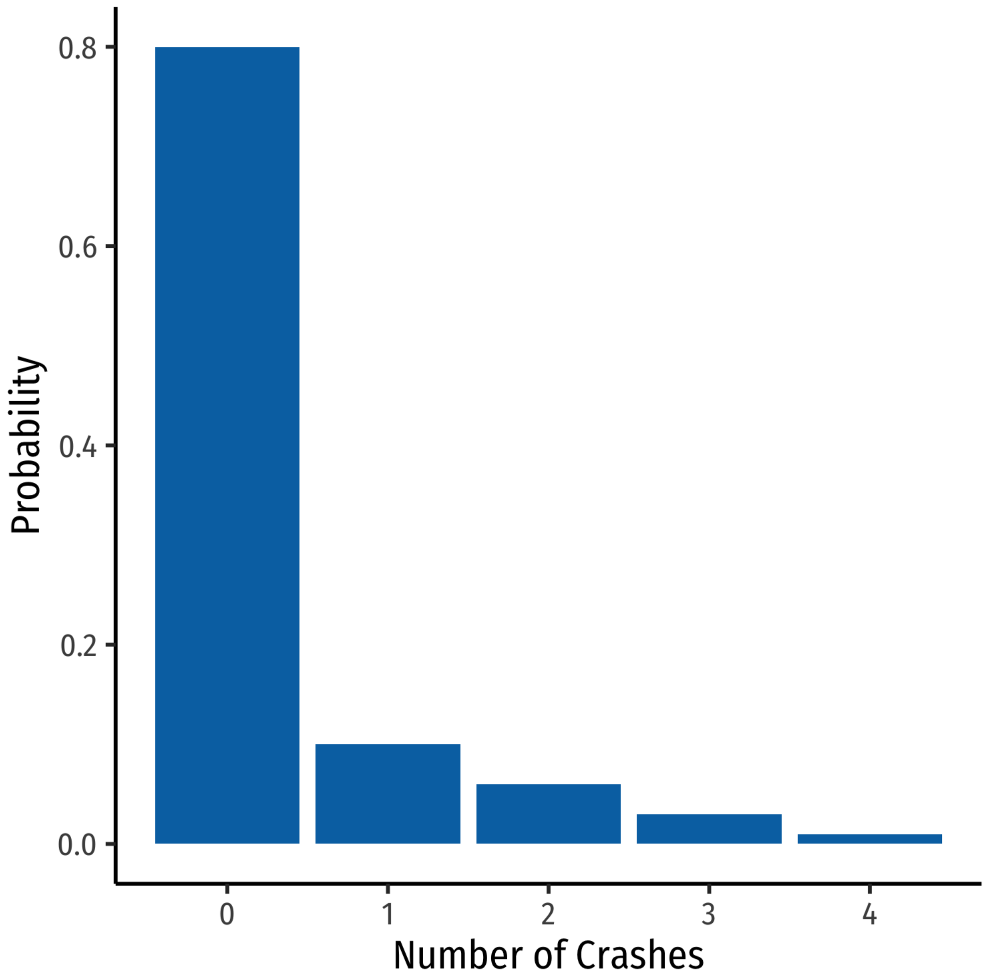
Example:

x_i	$P(X = x_i)$
0	0.80
1	0.10
2	0.06
3	0.03
4	0.01

Discrete Random Variables: pdf Graph

```r
crashes<-tibble(number = c(0,1,2,3,4),
                prob = c(0.80, 0.10, 0.06, 0.03, 0.01))

ggplot(data = crashes)+
  aes(x = number,
      y = prob)+
  geom_col(fill="#0072B2")+
  labs(x = "Number of Crashes",
       y = "Probability")+
    theme_classic(base_family = "Fira Sans Condensed",
           base_size=20)
```



Discrete Random Variables: cdf

Cumulative distribution function (pdf) lists probability X will be *at most* (less than or equal to) a given value x_i

- Notation: $F_X = P(X \leq x_i)$

Example:

x_i	$f(x)$	$F(x)$
0	0.80	0.80
1	0.10	0.90
2	0.06	0.96
3	0.03	0.99
4	0.01	1.00

- What is the probability your computer will crash *at most* once, $F(1)$?
- What about three times, $F(3)$?

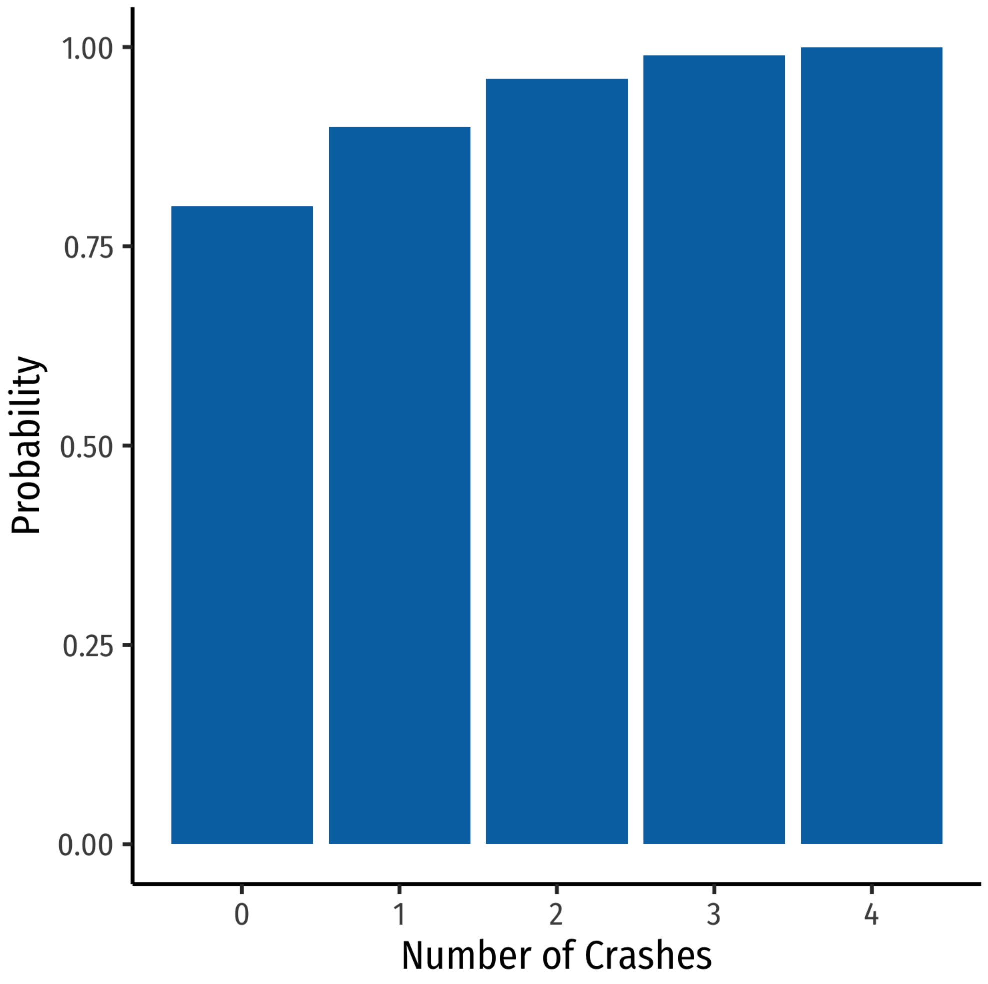
Discrete Random Variables: cdf Graph

```
crashes<-crashes %>%
  mutate(cum_prob = cumsum(prob))

crashes
```

```
## # A tibble: 5 × 3
##   number  prob cum_prob
##    <dbl> <dbl>    <dbl>
## 1      0  0.8      0.8
## 2      1  0.1      0.9
## 3      2  0.06     0.96
## 4      3  0.03     0.99
## 5      4  0.01     1
```

```
ggplot(data = crashes)+
  aes(x = number,
      y = cum_prob)+
  geom_col(fill="#0072B2")+
  labs(x = "Number of Crashes",
       y = "Probability")+
    theme_classic(base_family = "Fira Sans Condensed",
           base_size=20)
```



Expected Value and Variance

Expected Value of a Random Variable

- **Expected value** of a random variable X, written $E(X)$ (and sometimes μ), is the long-run average value of X "expected" after many repetitions

$$E(X) = \sum_{i=1}^{k} p_i x_i$$

- $E(X) = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k$

- A **probability-weighted average** of X, with each x_i weighted by its associated probability p_i

- Also called the **"mean"** or **"expectation"** of X, always denoted either $E(X)$ or μ_X

Expected Value: Example I

Example: Suppose you lend your friend \$100 at 10% interest. If the loan is repaid, you receive \$110. You estimate that your friend is 99% likely to repay, but there is a default risk of 1% where you get nothing. What is the expected value of repayment?

Expected Value: Example II

Example:

Let X be a random variable that is described by the following pdf:

x_i	$P(X = x_i)$
1	0.50
2	0.25
3	0.15
4	0.10

Calculate $E(X)$.

The Steps to Calculate E(X), Coded

```r
# Make a Random Variable called X
X<-tibble(x_i=c(1,2,3,4), # values of X
          p_i=c(0.50,0.25,0.15,0.10)) # probabilities

X %>%
  summarize(expected_value = sum(x_i*p_i))
```

```
## # A tibble: 1 × 1
##   expected_value
##            <dbl>
## 1           1.85
```

Variance of a Random Variable

- The **variance** of a random variable X, denoted $var(X)$ or σ_X^2 is:

$$\begin{aligned}\sigma_X^2 &= E[(x_i-\mu_X)^2]\\ &= \sum_{i=1}^{n}(x_i-\mu_X)^2 p_i\end{aligned}$$

- This is the **expected value of the squared deviations from the mean**
    - i.e. the probability-weighted average of the squared deviations

Standard Deviation of a Random Variable

- The **standard deviation** of a random variable X, denoted $sd(X)$ or σ_X is:

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

Standard Deviation: Example I

Example: What is the standard deviation of computer crashes?

x_i	$P(X = x_i)$
0	0.80
1	0.10
2	0.06
3	0.03
4	0.01

The Steps to Calculate sd(X), Coded I

```
# get the expected value
crashes %>%
  summarize(expected_value = sum(number*prob))
```

```
## # A tibble: 1 × 1
##   expected_value
##            <dbl>
## 1           0.35
```

```
# save this for quick use
exp_value<-0.35

crashes_2 <- crashes %>%
  select(-cum_prob) %>% # we don't need the cdf
  # create new columns
  mutate(deviations = number - exp_value, # deviations from exp_value
         deviations_sq = deviations^2,
         weighted_devs_sq = prob * deviations^2) # square deviations
```

The Steps to Calculate sd(X), Coded II

```r
# look at what we made
crashes_2
```

```
## # A tibble: 5 × 5
##   number  prob deviations deviations_sq weighted_devs_sq
##    <dbl> <dbl>      <dbl>         <dbl>            <dbl>
## 1      0  0.8       -0.35         0.122           0.098
## 2      1  0.1        0.65         0.423           0.0423
## 3      2  0.06       1.65         2.72            0.163
## 4      3  0.03       2.65         7.02            0.211
## 5      4  0.01       3.65        13.3             0.133
```

The Steps to Calculate sd(X), Coded III

```r
# now we want to take the expected value of the squared deviations to get variance
crashes_2 %>%
  summarize(variance = sum(weighted_devs_sq), # variance
            sd = sqrt(variance)) # sd is square root
```

```
## # A tibble: 1 × 2
##   variance    sd
##      <dbl> <dbl>
## 1    0.648 0.805
```

Standard Deviation: Example II

Example: What is the standard deviation of the random variable we saw before?

x_i	$P(X = x_i)$
1	0.50
2	0.25
3	0.15
4	0.10

Hint: you already found it's expected value.

Continuous Random Variables

Continuous Random Variables

- **Continuous random variables** can take on an uncountable (infinite) number of values

- So many values that the probability of any specific value is infinitely small:

$$P(X = x_i) \rightarrow 0$$

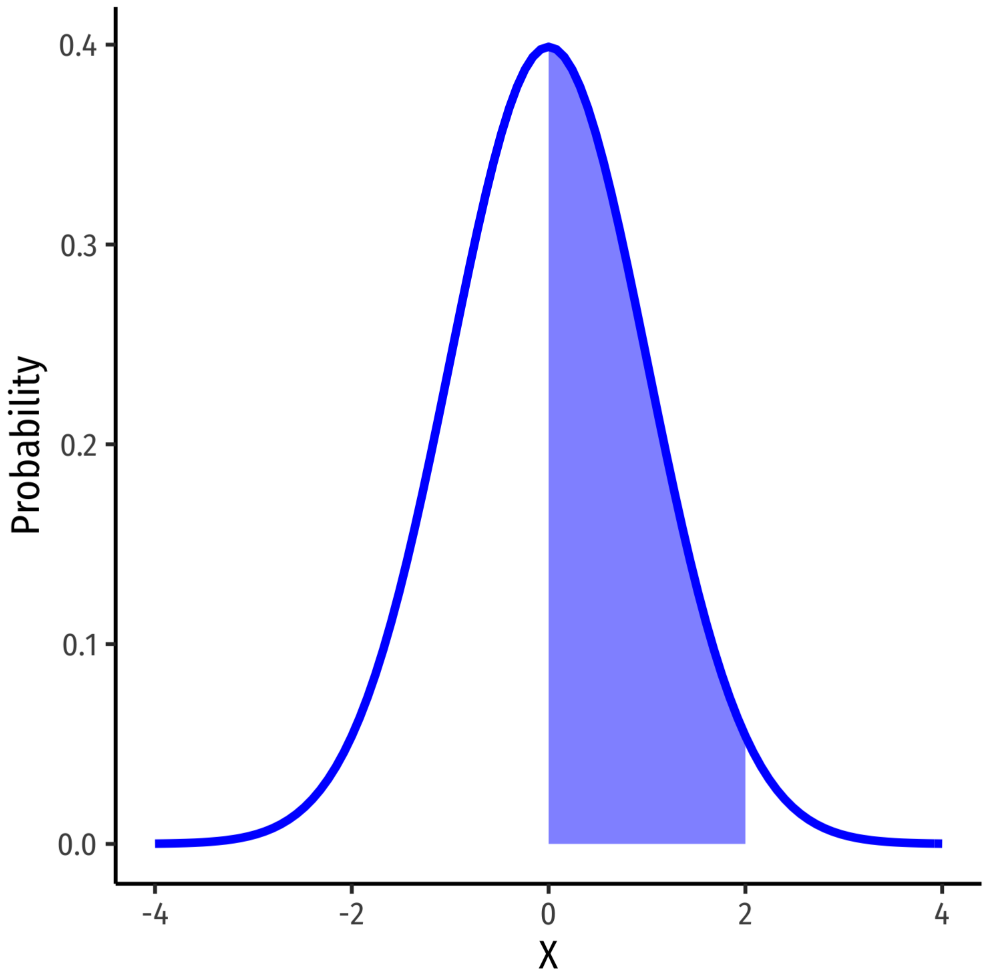
- Instead, we focus on a *range* of values it might take on

Continuous Random Variables: pdf I

Probability *density* function (pdf) of a continuous variable represents the probability between two values as the area under a curve

- The total area under the curve is 1
- Since $P(a) = 0$ and $P(b) = 0$, $P(a < X < b) = P(a \leq X \leq b)$

Example: $P(0 \leq X \leq 2)$

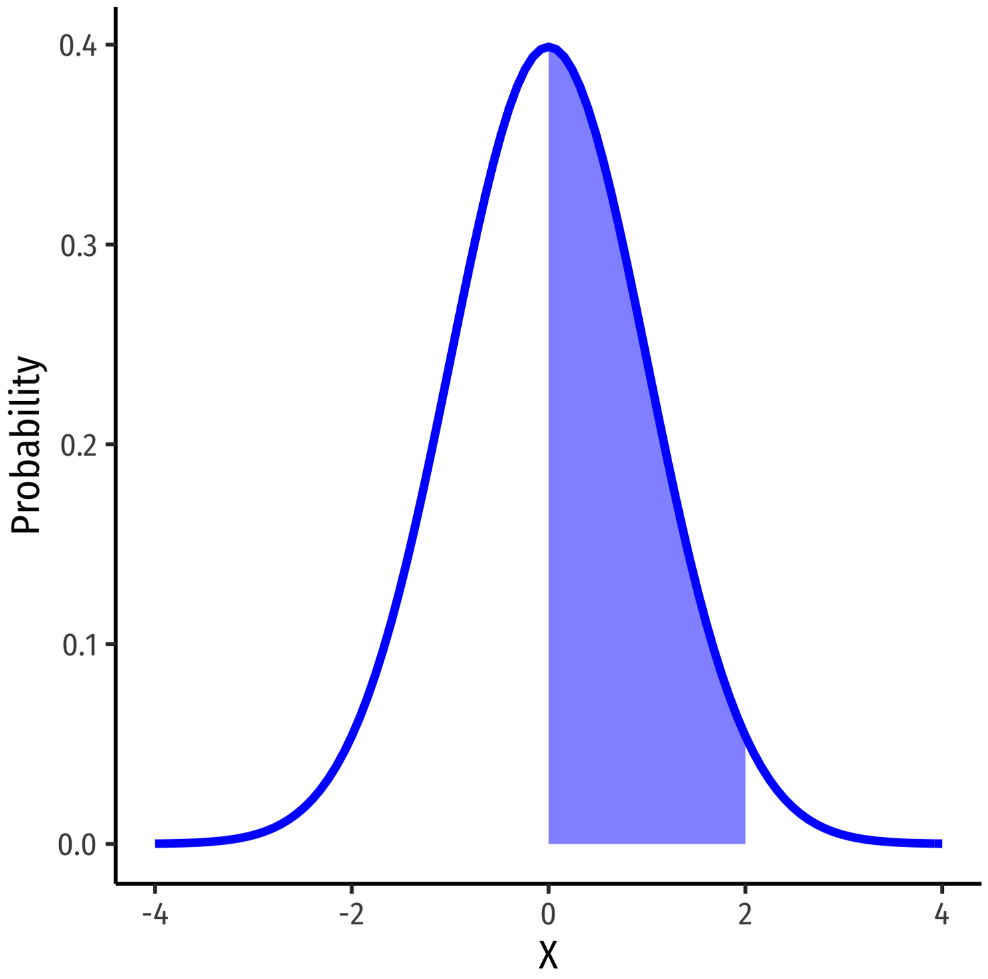


Continuous Random Variables: pdf II

- FYI using calculus:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Complicated: software or (old fashioned!) probability tables to calculate

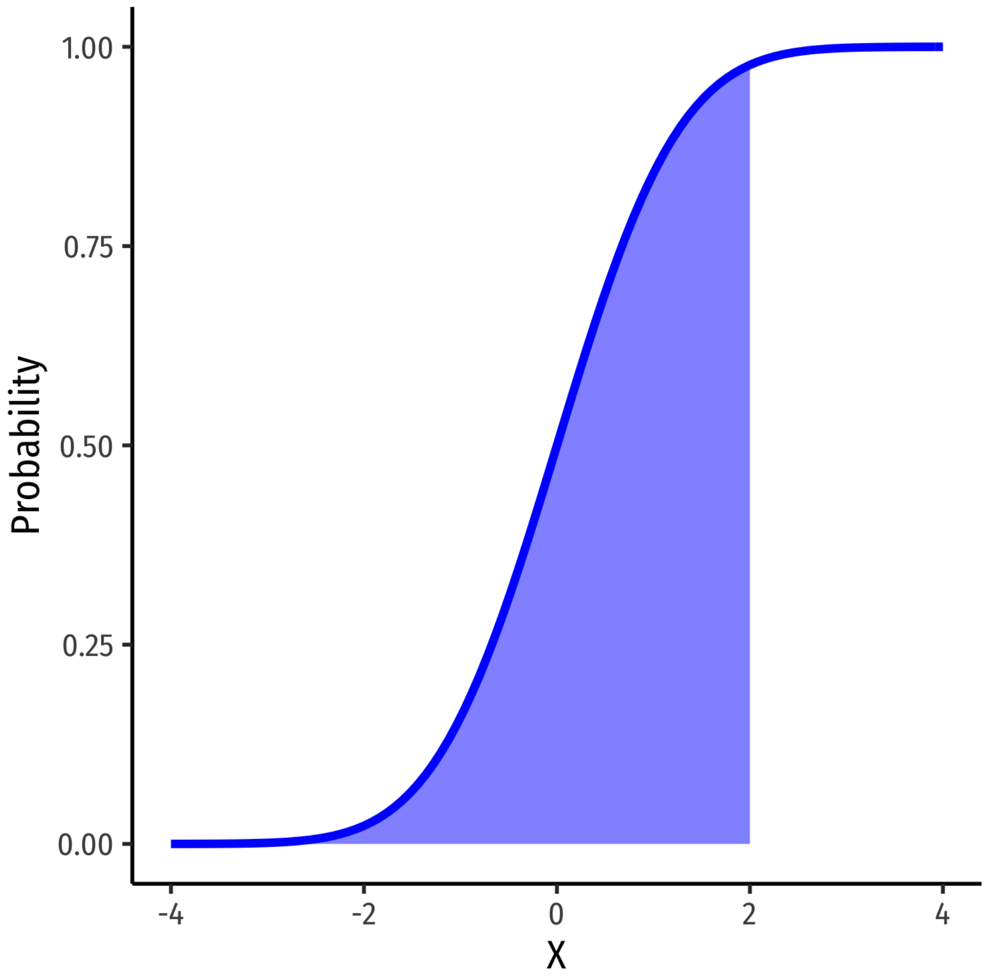


Continuous Random Variables: cdf I

- The **cumulative density function (cdf)** describes the area under the pdf for all values less than or equal to (i.e. to the left of) a given value, k

$$P(X \leq k)$$

Example: $P(X \leq 2)$



Continuous Random Variables: cdf II

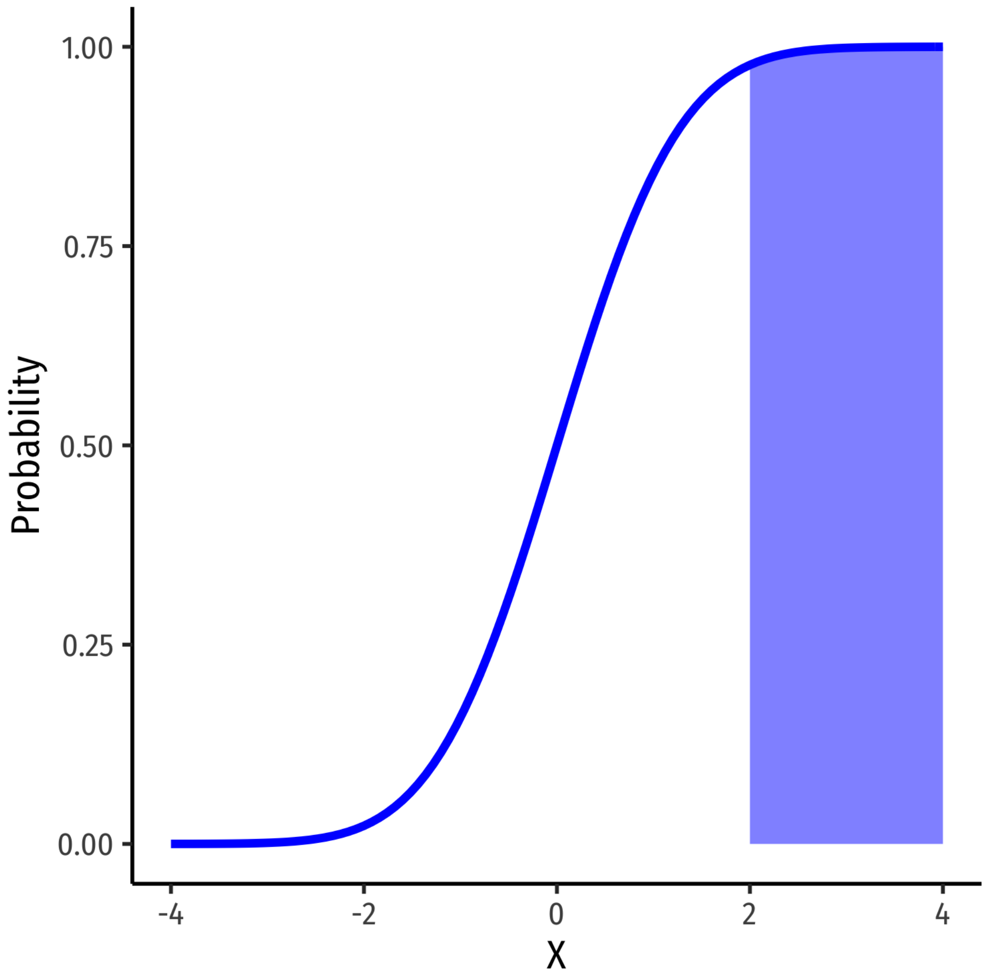
- Note: to find the probability of values *greater* than or equal to (to the right of) a given value k:

$$P(X \geq k) = 1 - P(X \leq k)$$

Example:

$P(X \geq 2) = 1 - P(X \leq 2)$

$P(X \geq 2) =$ area under the curve to the right of 2



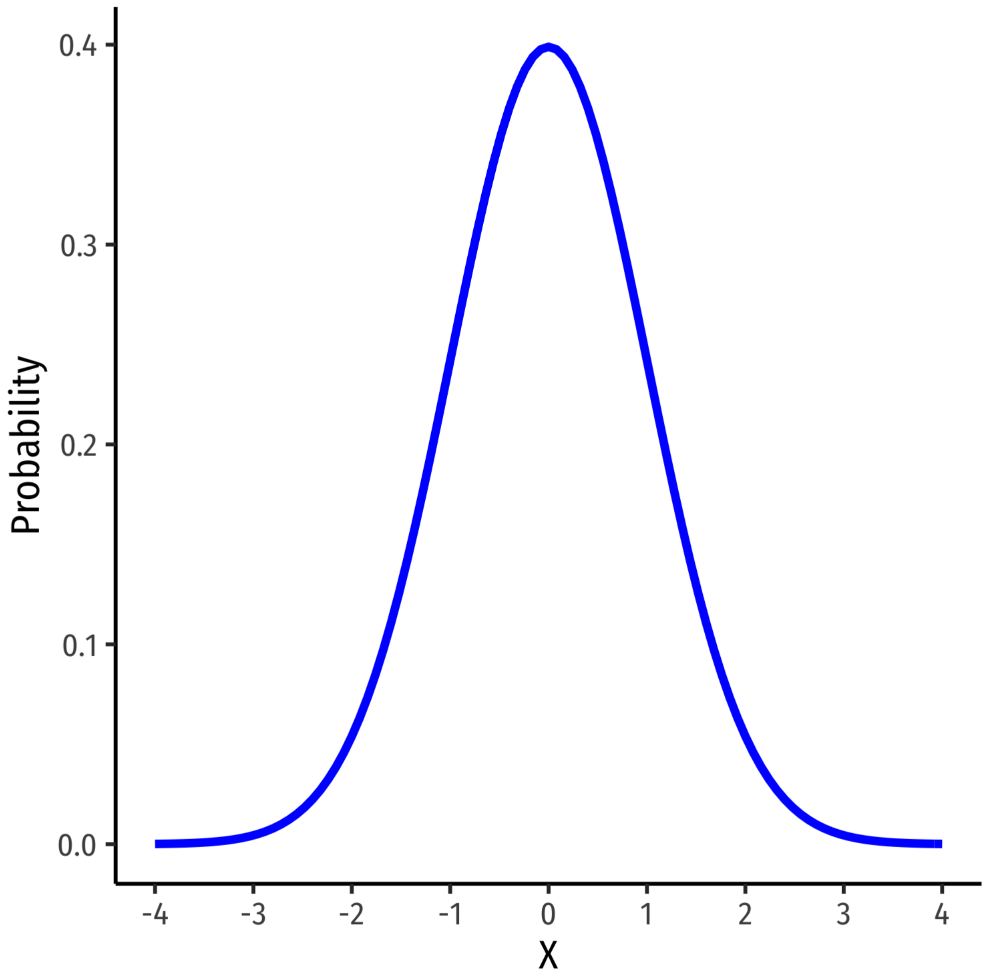
The Normal Distribution

The Normal Distribution I

- The **Gaussian** or **normal distribution** is the most useful type of probability distribution

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

- Continuous, symmetric, unimodal, with mean μ and standard deviation σ

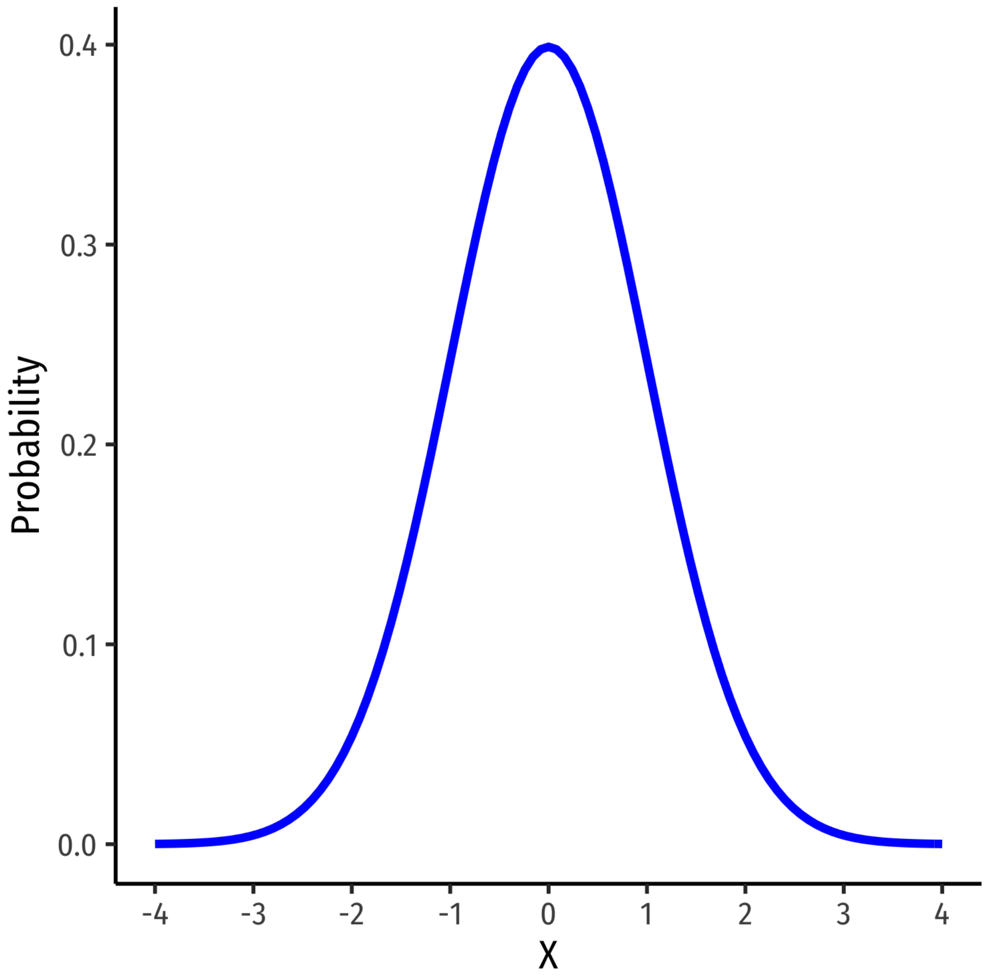


The Normal Distribution II

- FYI: The pdf of $X \sim N(\mu, \sigma)$ is

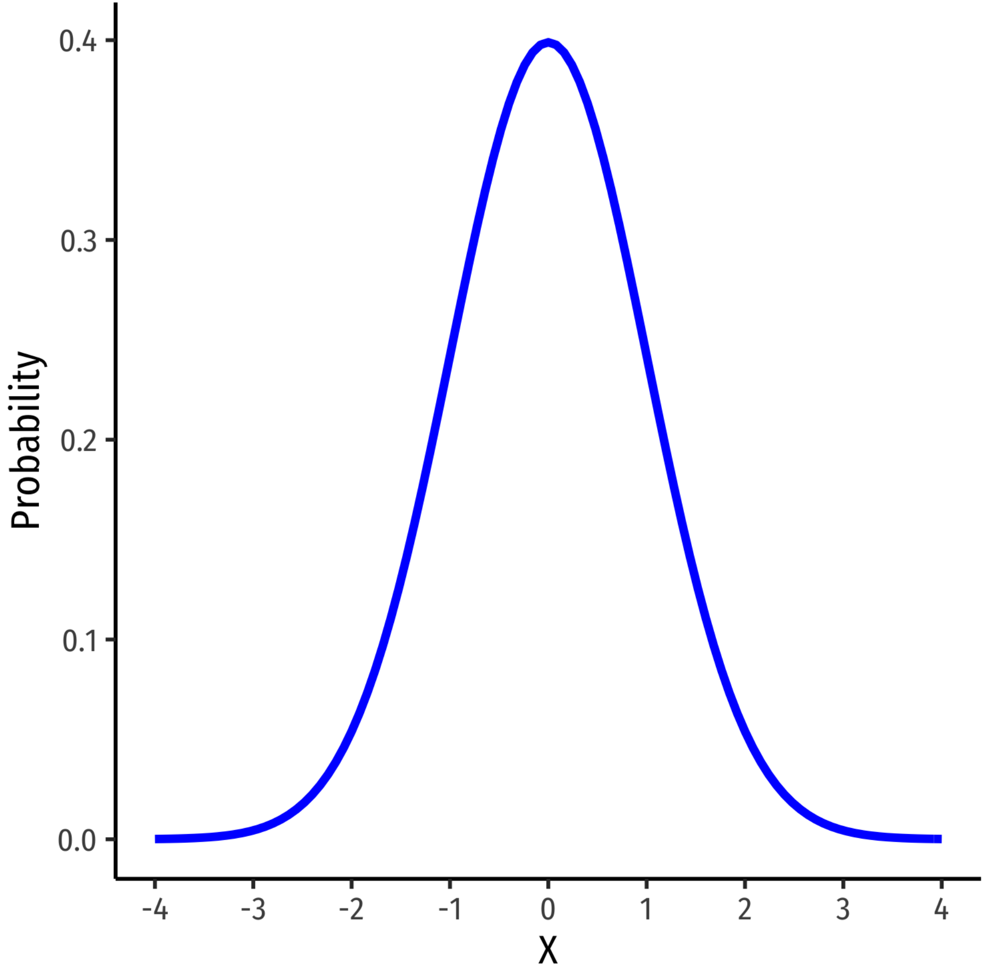
$$P(X=k)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{(k-\mu)}{\sigma}\right)^2}$$

- **Do not try and learn this**, we have software and (previously tables) to calculate pdfs and cdfs



The 68-95-99.7 Rule

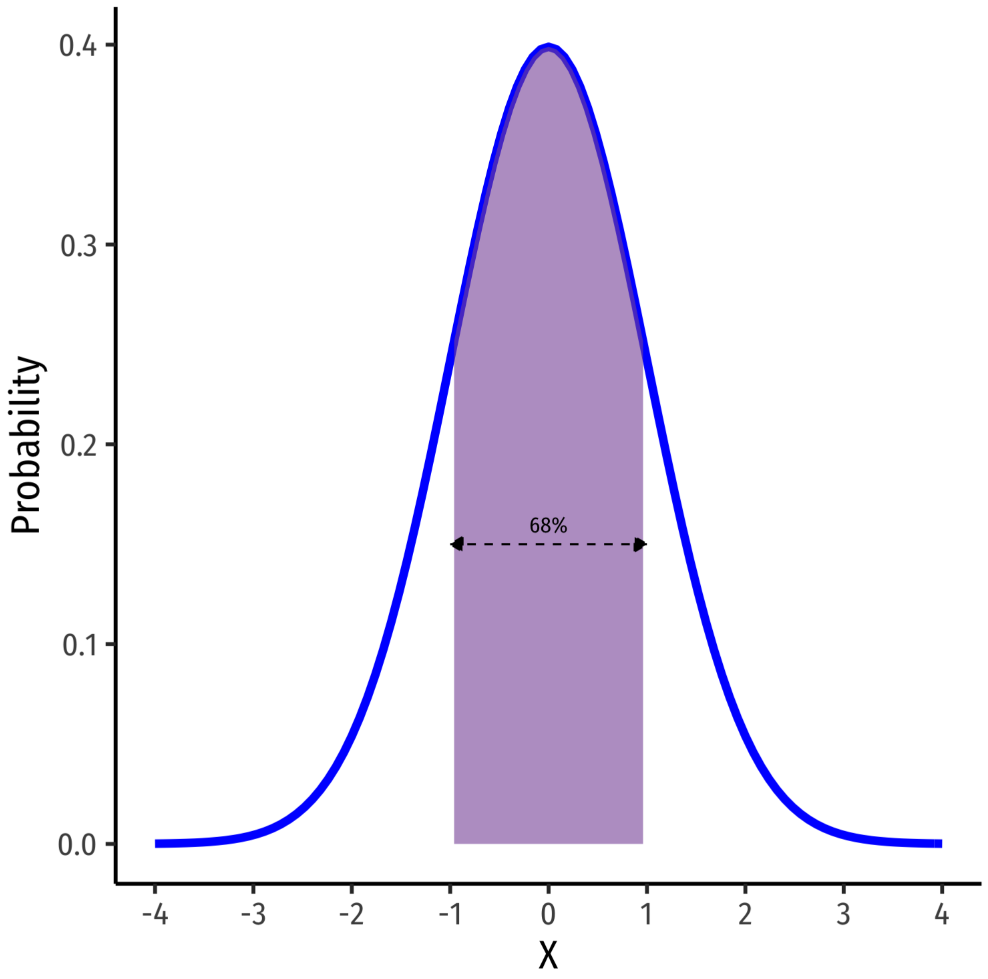
- **68-95-99.7% empirical rule**: for a normal distribution:



The 68-95-99.7 Rule

- **68-95-99.7% empirical rule**: for a normal distribution:

- $P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx$ 68%

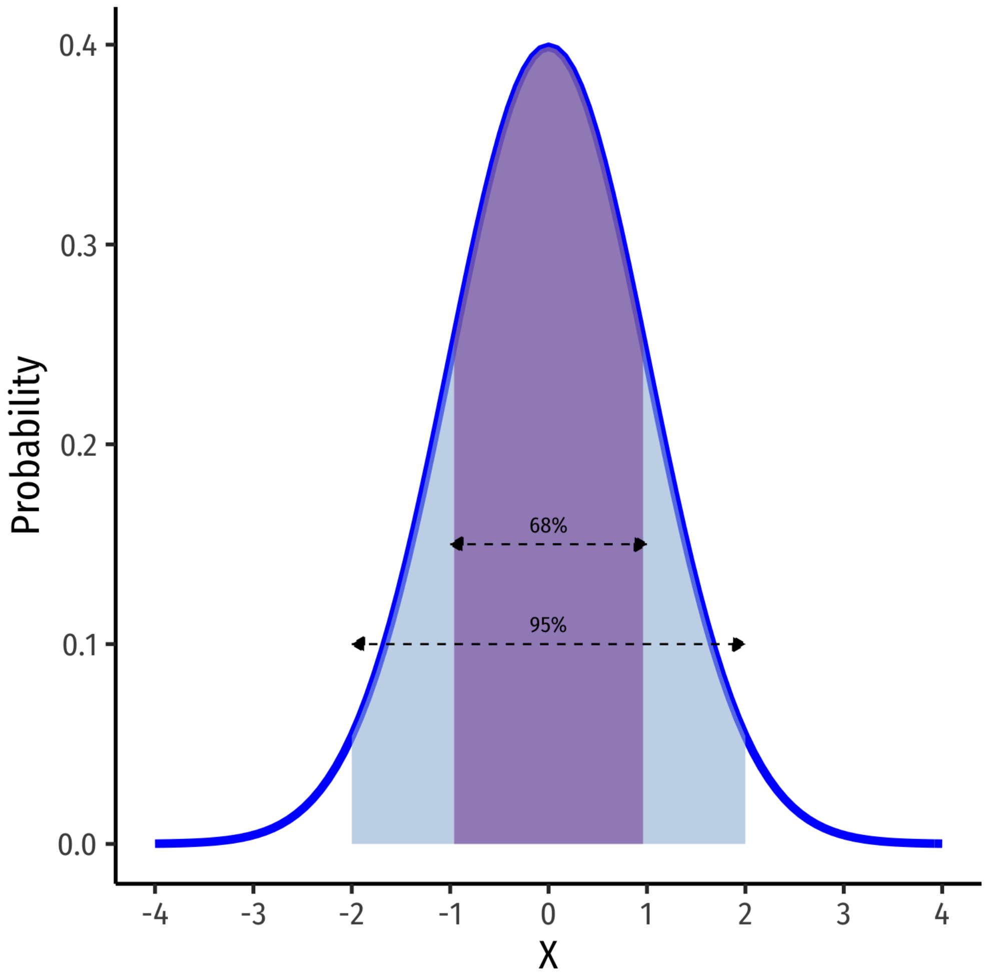


The 68-95-99.7 Rule

- **68-95-99.7% empirical rule**: for a normal distribution:

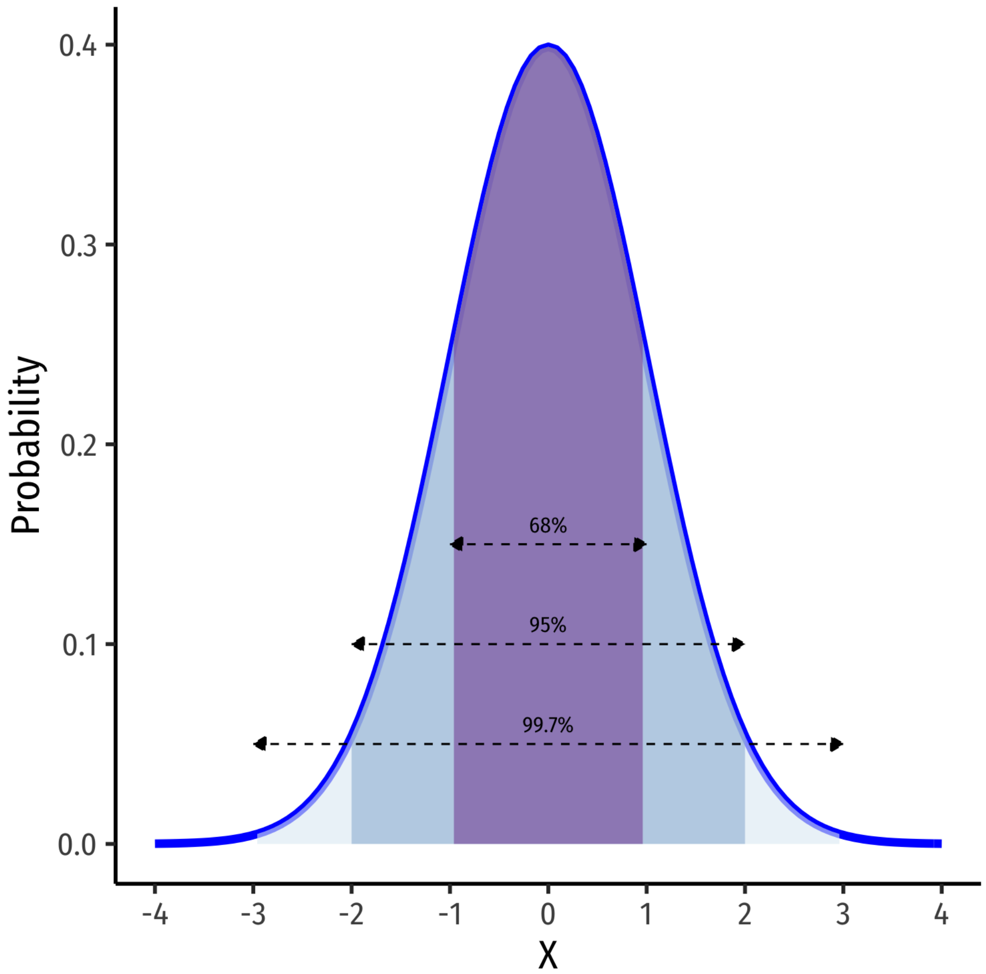
- $P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx$ 68%

- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx$ 95%



The 68-95-99.7 Rule

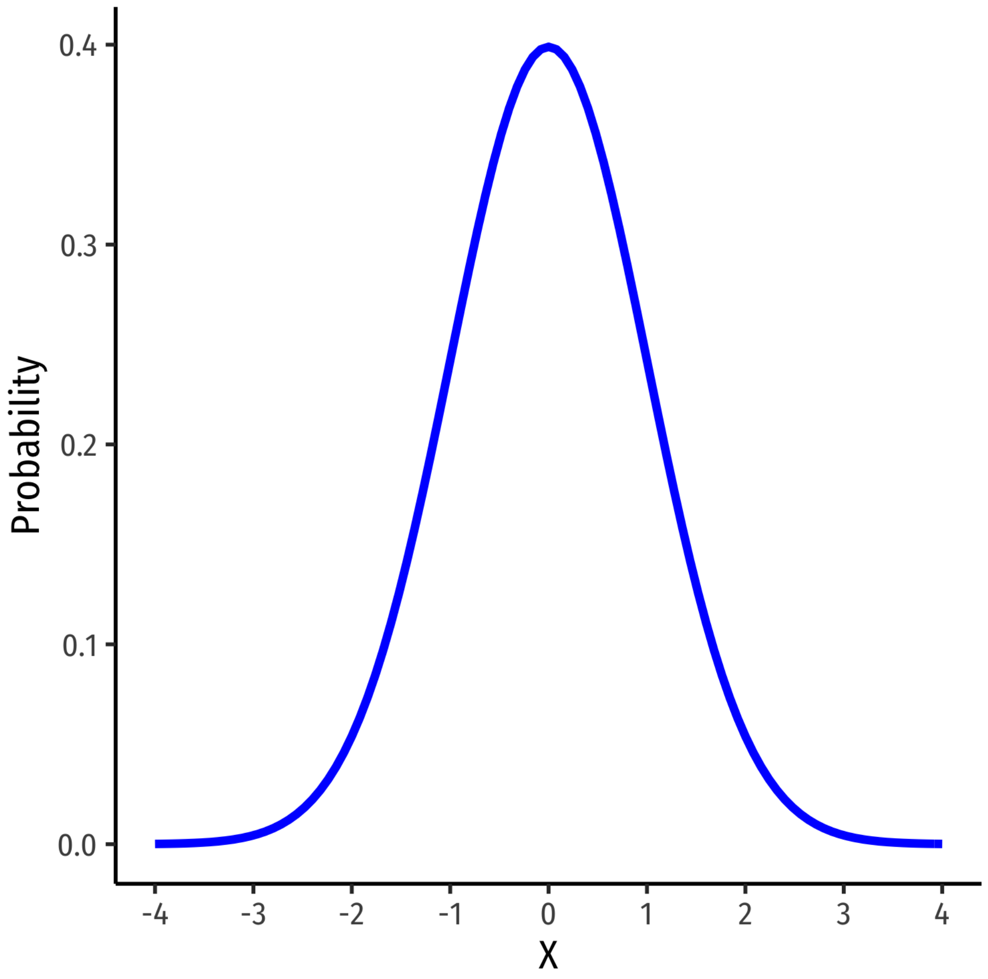
- **68-95-99.7% empirical rule**: for a normal distribution:
- $P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx$ 68%
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx$ 95%
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx$ 99.7%
- **68/95/99.7%** of observations fall within **1/2/3 standard deviations** of the mean



The Standard Normal Distribution

- The **standard** normal distribution (often referred to as **Z**) has mean 0 and standard deviation 1

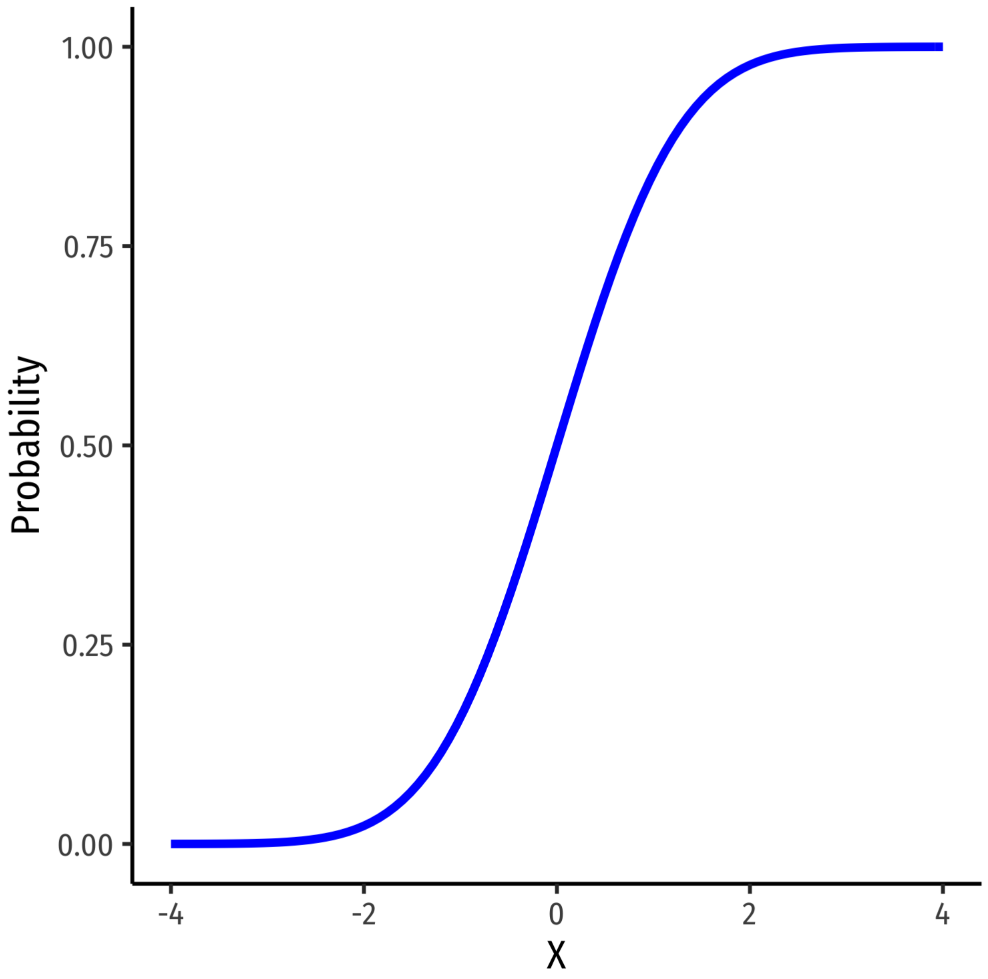
$$Z \sim N(0, 1)$$



The Standard Normal cdf

- The **standard** normal cdf

$$\Phi(k) = P(Z \leq k)$$



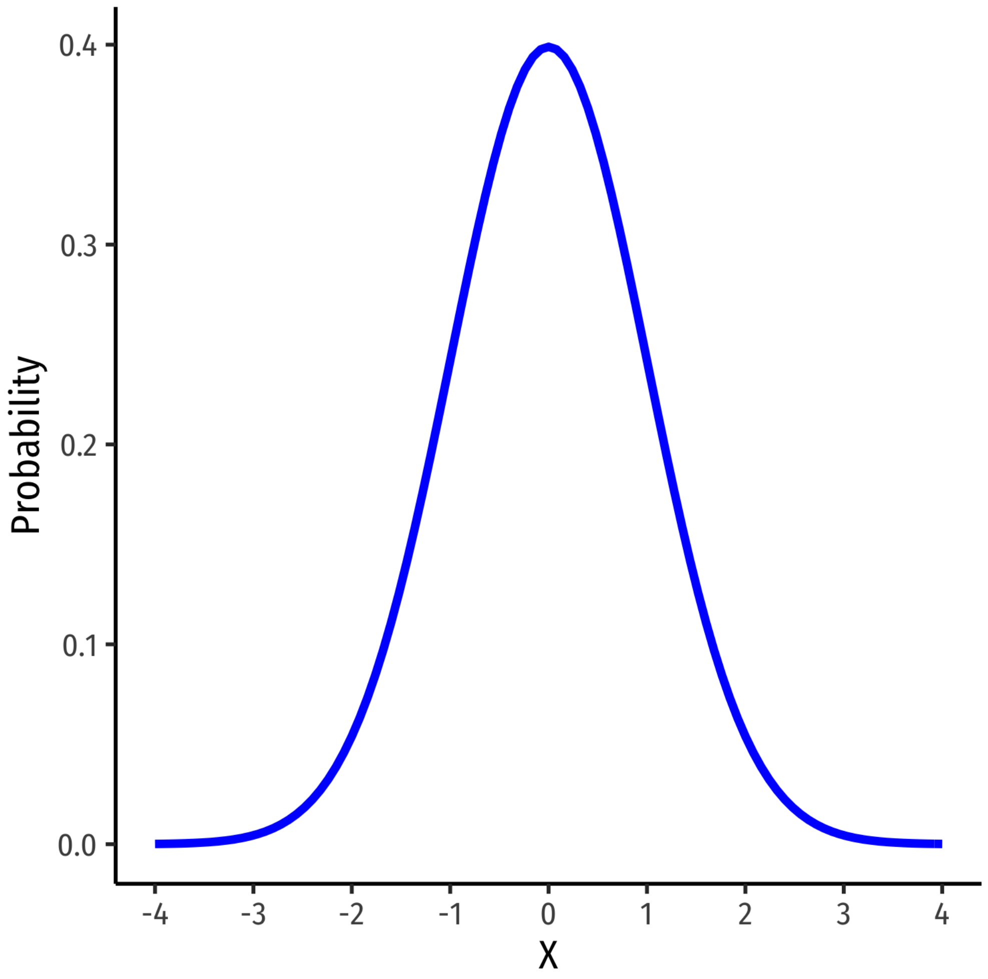
Standardizing Variables

- We can take any normal distribution (for any μ, σ) and **standardize** it to the standard normal distribution by taking the **Z-score** of any value, x_i:

$$Z = \frac{x_i - \mu}{\sigma}$$

- Subtract any value by the distribution's mean and divide by standard deviation

- Z: number of standard deviations x_i value is away from the mean



Standardizing Variables: Example

Example: On August 8, 2011, the Dow dropped 634.8 points, sending shock waves through the financial community. Assume that during mid-2011 to mid-2012 the daily change for the Dow is normally distributed, with the mean daily change of 1.87 points and a standard deviation of 155.28 points. What is the Z-score?

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{634.8 - 1.87}{155.28}$$

$$Z = -4.1$$

This is 4.1 standard deviations (σ) beneath the mean, an *extremely* low probability event.

Standardizing Variables: From X to Z I

Example: In the last quarter of 2015, a group of 64 mutual funds had a mean return of 2.4% with a standard deviation of 5.6%. These returns can be approximated by a normal distribution.

What percent of the funds would you expect to be earning between -3.2% and 8.0% returns?

Convert to standard normal to find Z-scores for 8 and -3.2.

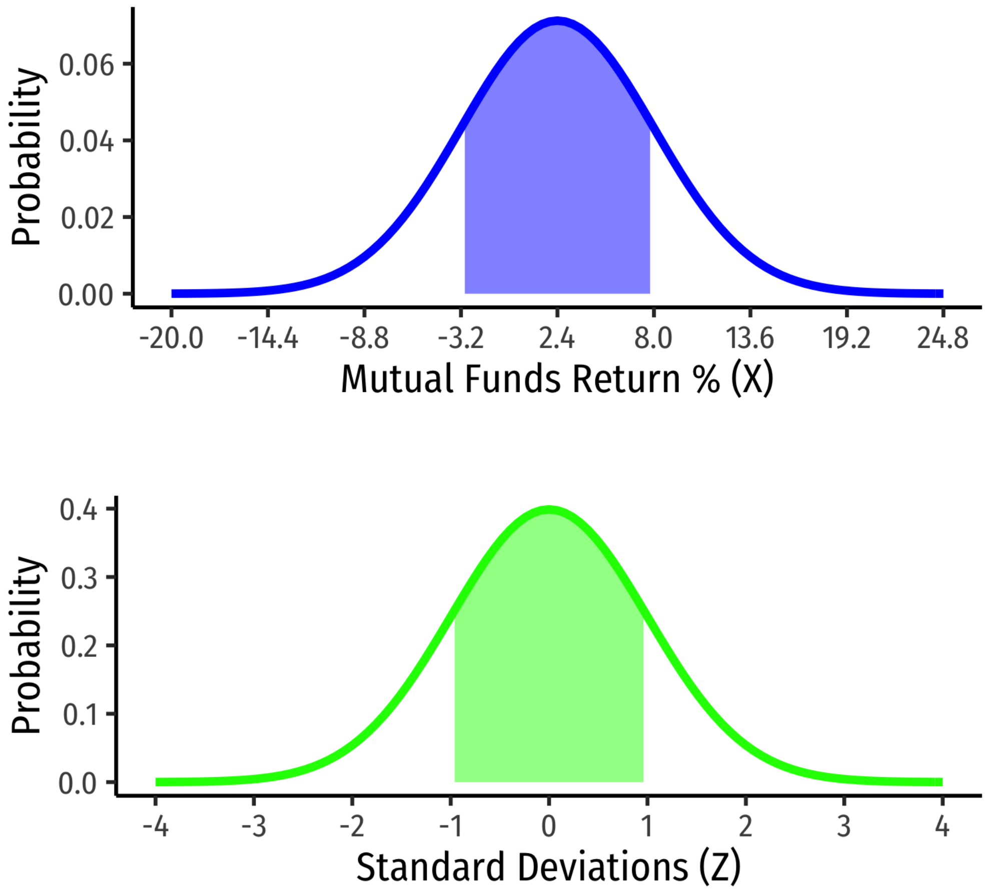
$$P(-3.2 < X < 8)$$

$$P(\frac{-3.2 - 2.4}{5.6} < \frac{X - 2.4}{5.6} < \frac{8 - 2.4}{5.6})$$

$$P(-1 < Z < 1)$$

$$P(X \pm 1\sigma) = 0.68$$

Standardizing Variables: From X to Z II



Standardizing Variables: From X to Z III

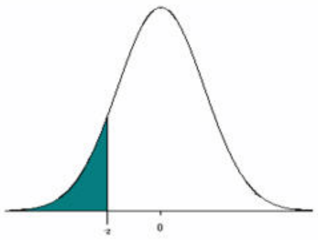
You Try: In the last quarter of 2015, a group of 64 mutual funds had a mean return of 2.4% with a standard deviation of 5.6%. These returns can be approximated by a normal distribution.

1. What percent of the funds would you expect to be earning between -3.2% and 8.0% returns?
2. What percent of the funds would you expect to be earning 2.4% or less?
3. What percent of the funds would you expect to be earning between -8.8% and 13.6%?
4. What percent of the funds would you expect to be earning returns greater than 13.6%?

Finding Z-score Probabilities I

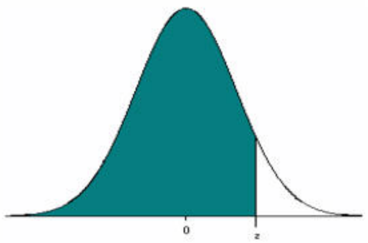
- How do we actually find the probabilities for $Z-$scores?

Table of Standard Normal Probabilities for Negative Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681

Table of Standard Normal Probabilities for Positive Z-scores

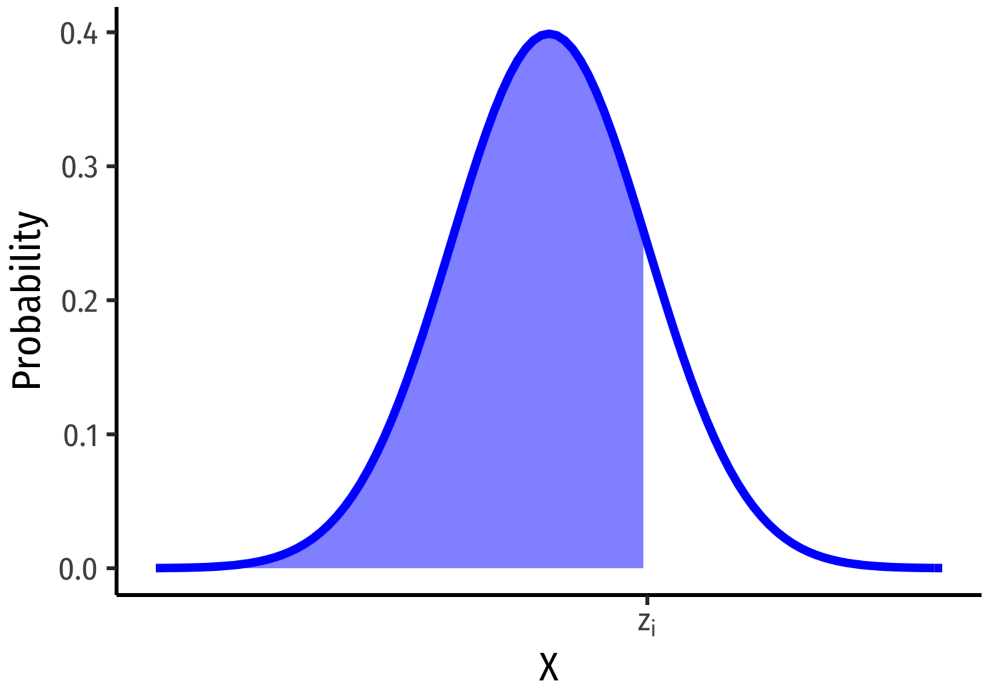


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Finding Z-score Probabilities II

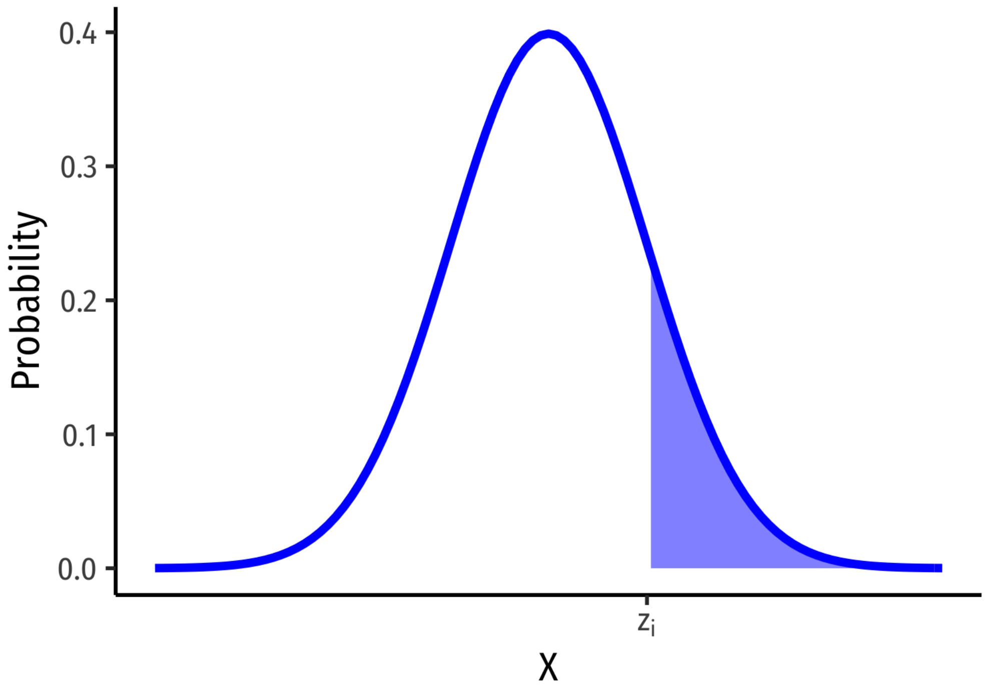
Probability to the **left** of z_i

$$P(Z \leq z_i) = \underbrace{\Phi(z_i)}_{\text{cdf of } z_i}$$



Probability to the **right** of z_i

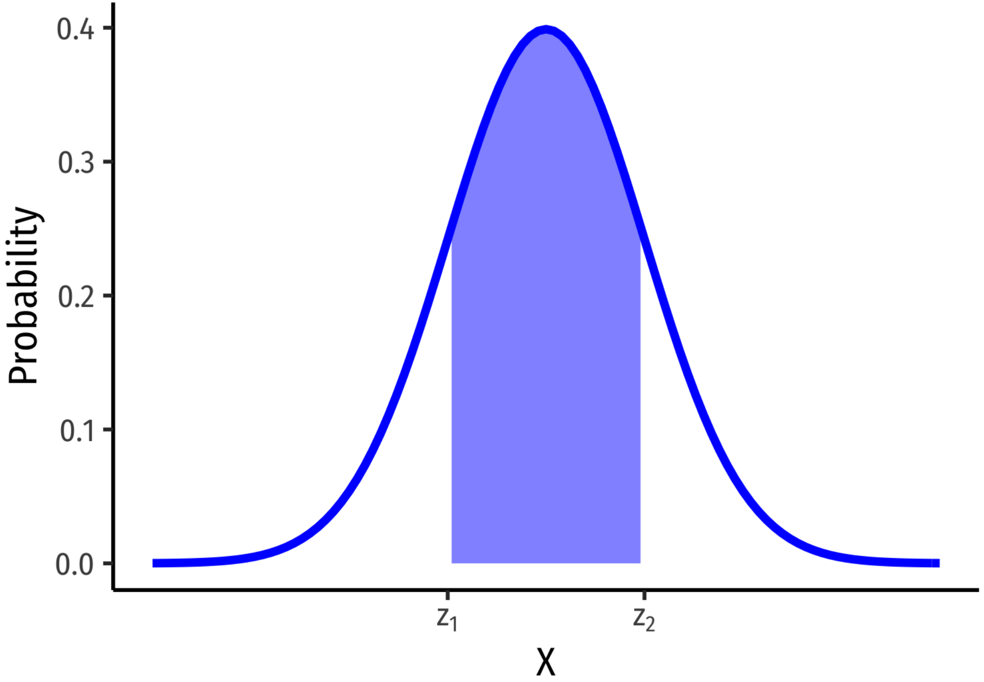
$$P(Z \geq z_i) = 1 - \underbrace{\Phi(z_i)}_{\text{cdf of } z_i}$$



Finding Z-score Probabilities III

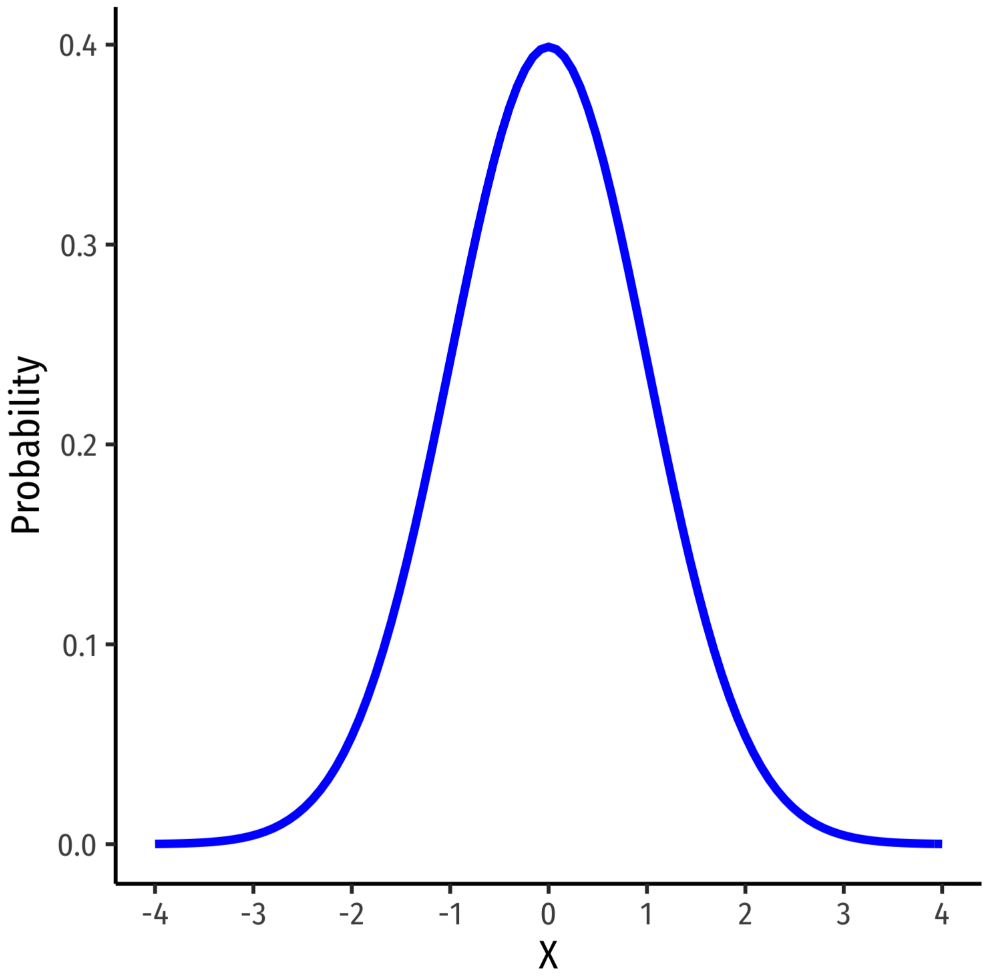
Probability **between** z_1 and z_2

$$P(z_1 \geq Z \geq z_2) = \underbrace{\Phi(z_2)}_{\text{cdf of } z_2} - \underbrace{\Phi(z_1)}_{\text{cdf of } z_1}$$



Finding Z-score Probabilities IV

- `pnorm()` calculates `p`robabilities with a `norm`al distribution with arguments:
  - `mean =` the mean
  - `sd =` the standard deviation
  - `lower.tail =`
    - `TRUE` if looking at area to *LEFT* of value
    - `FALSE` if looking at area to *RIGHT* of value



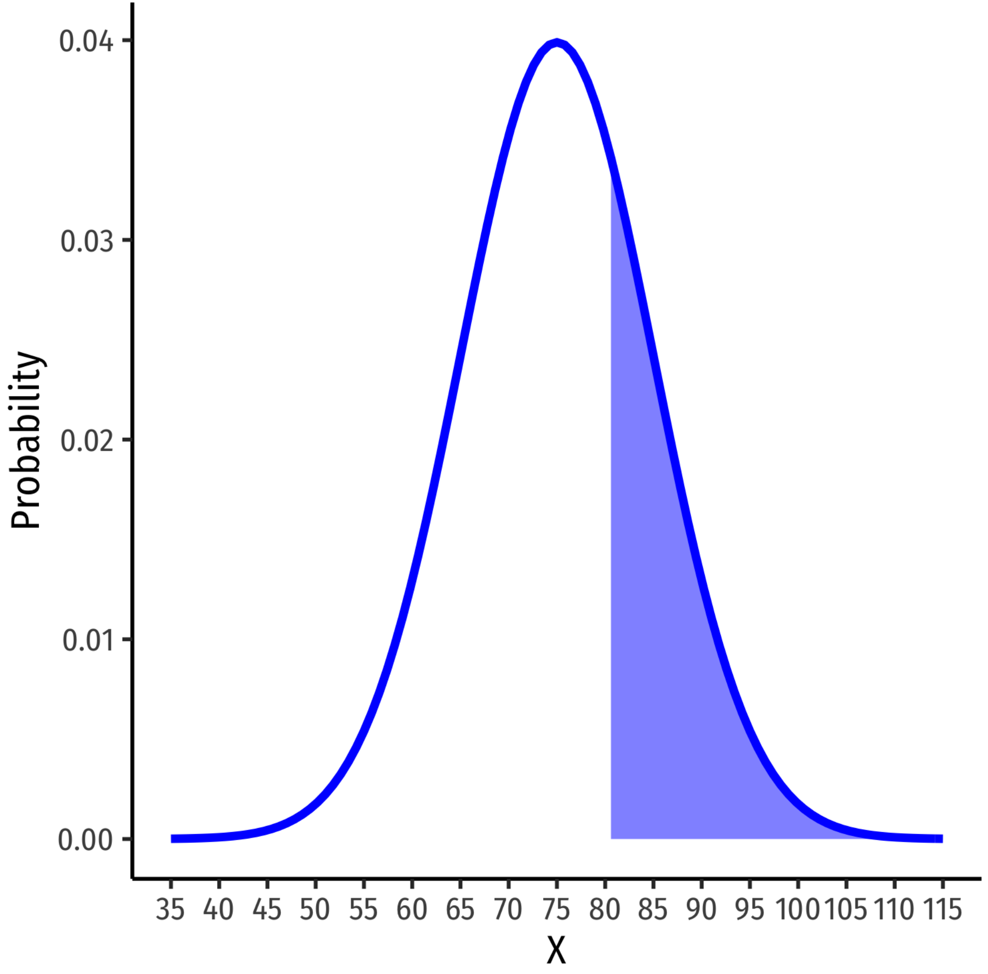
Finding Z-score Probabilities IV

Example: Let the distribution of grades be normal, with mean 75 and standard deviation 10.

- Probability a student gets **at least an 80**

```
pnorm(80,
      mean = 75,
      sd = 10,
      lower.tail = FALSE) # looking to right
```

```
## [1] 0.3085375
```



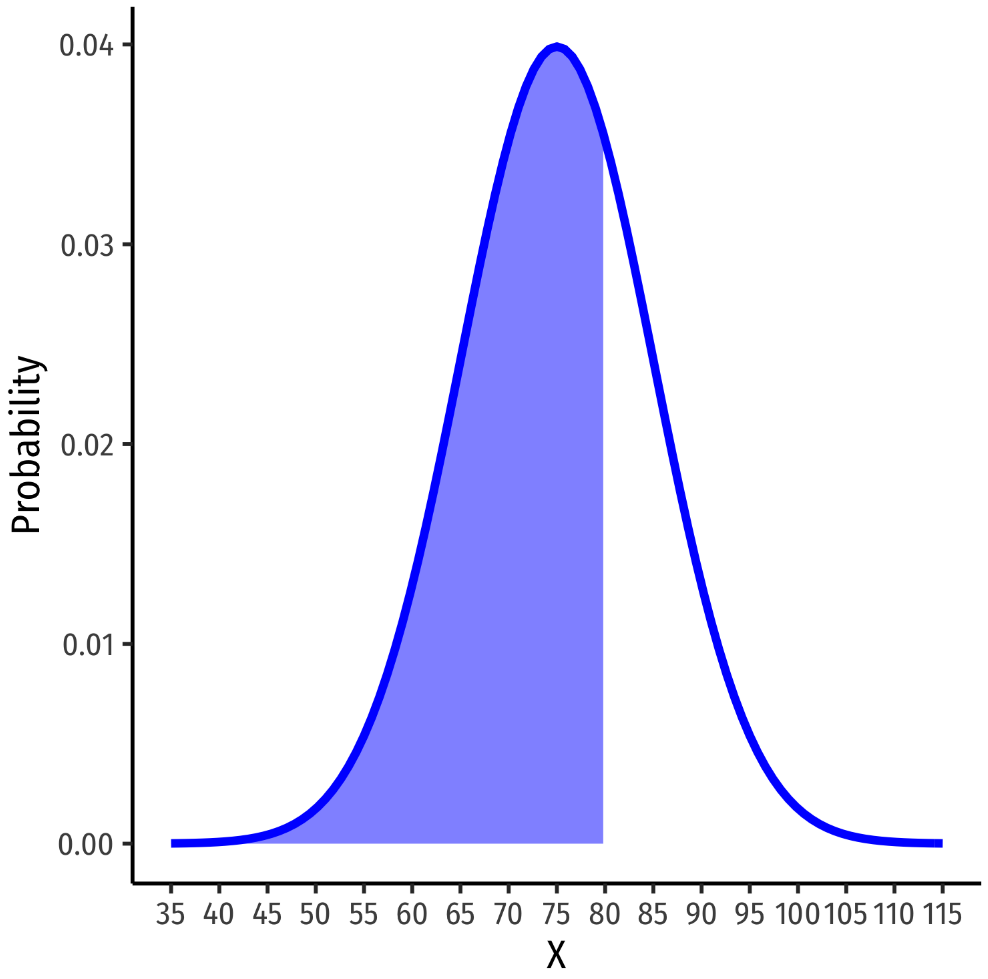
Finding Z-score Probabilities V

Example: Let the distribution of grades be normal, with mean 75 and standard deviation 10.

- Probability a student gets **at most an 80**

```
pnorm(80,
      mean = 75,
      sd = 10,
      lower.tail = TRUE) # looking to left
```

```
## [1] 0.6914625
```



Finding Z-score Probabilities VI

Example: Let the distribution of grades be normal, with mean 75 and standard deviation 10.

- Probability a student gets **between a 65 and 85**

```
# subtract two left tails!
pnorm(85, # larger number first!
      mean = 75,
      sd = 10,
      lower.tail = TRUE) - # looking to left, & SUBTRACT
  pnorm(65, # smaller number second!
        mean = 75,
        sd = 10,
        lower.tail = TRUE) #looking to left
```

```
## [1] 0.6826895
```