Probability review

Sample space - the set of all possible outcomes

eg: roll a die, sample space: $S = \{1, 2, 3, 4, 5, 6\}$

An event is a subset of the sample space

eg: A: roll a three $A = \{3\} \subset S = \{1, 2, 3, 4, 5, 6\}$

B: roll an even $B = \{2, 4, 6\} \subset S$

C: roll something $C = S \subset S$

$A \cup B$ = A or B — roll a three or an even = $\{2, 3, 4, 6\}$

$A \cap B$ = A and B — roll a three and an even = $\{\} = \emptyset$

empty set / non-event

complement: $\bar{A}$ = not A — not rolling a three = $\{1, 2, 4, 5, 6\}$

If $A \cap B = \emptyset$ then A and B are mutually exclusive or disjoint events.

For a discrete (i.e. finite or countable) sample space S a probability function is a function $f : S \to \mathbb{R}$ such that

- for each outcome $x \in S$, $0 \leq f(x) \leq 1$
- $\sum_{x \in S} f(x) = 1$.

eg: probability function for rolling a fair die is $f(x) = \frac{1}{6}$

Given a probability function $f : S \to \mathbb{R}$, the probability of an event $A \subset S$ is

$$P(A) = \sum_{x \in A} f(x)$$

eg: probability of rolling an even $= \sum_{x \in \{2,4,6\}} f(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

Properties:

$P(\phi) = 0$

$P(\bar{A}) = 1 - P(A)$

$P(\bar{A} \cap B) = P(B) - P(A \cap B)$

$P(A \cup B) = P(A) + P(B) + P(A \cap B)$

The conditional probability of an event A occurring given that B has occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{when } P(B) \neq 0$$

$$\text{else } P(A \mid B) = 0$$

eg: roll a 2 (A) given an even has been rolled (B)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

We say A is independent of B if $P(A \mid B) = P(A)$

In which case $P(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

$$\Rightarrow P(A \cap B) = P(A)\,P(B)$$

eg: toss a coin twice, A = second toss tails

B = first toss tails

A is independent of B, so $P(A \mid B) = P(A) = \frac{1}{2}$

and the probability of consecutive tails

$$P(A \cap B) = P(A)\,P(B) = \frac{1}{4}$$

Discrete random variables

A random variable X is a function $X : S \to \mathbb{R}$ from a sample space to the reals.

If we have a probability function then we denote

$$P(X(s) = x) = P(\{s \in S \mid X(s) = x\})$$

$x \in \mathbb{R}$ $\quad P(X = x) =$ probability of a value $x \in \mathbb{R}$ for the random variable $X : S \to \mathbb{R}$

Example

roll a blue die and a green die

Sample space $S = \{(1,1), (1,2), (2,1), \ldots, (6,6)\}$

$6 \times 6 = 36$ possible outcomes

define a random variable

$$X : S \to \mathbb{R}$$

$$(a, b) \mapsto a + b \qquad \text{range } X = \{2, 3, \ldots, 12\}$$

$P(X = 2) = \frac{1}{36}$ (only one way to roll: $(1,1)$)

$P(X = 3) = \frac{2}{36} = \frac{1}{18}$ $\quad (1,2), (2,1)$

$P(X = 7) = \frac{6}{36} = \frac{1}{6}$

The above is an example of a discrete random variable because the range of $X : S \to \mathbb{R}$ is a discrete subset of $\mathbb{R}$.

The probability mass function (pmf) for a discrete random variable $X : S \to \mathbb{R}$ is the function $p_X : \mathbb{R} \to \mathbb{R}$

$$p_X(x) = P(X = x) = P(\{s \in S \mid X(s) = x\})$$

From the axioms defining a probability function it follows that

$$0 \le p_X(x) \le 1 \quad \text{and} \quad \sum_{x \in \text{range}(X)} p_X(x) = 1$$

eg: pmf for the green and blue dice example

x	2	3	...	12	otherwise
$p_X(x)$	$\frac{1}{36}$	$\frac{1}{18}$		$\frac{1}{36}$	0

The cumulative distribution function (cdf) for a discrete r.v. X is

$$F_X(t) = P(X \le t) = \sum_{x \le t} p_X(x)$$

The expected value $E[X]$ or μ_X is

$$E[X] = \sum_{\text{range of } X} x\, p_X(x) \qquad \text{(mean)}$$

eg: a friend proposes a game where he rolls a die and gives you $\$(4-s)$, where s is the number rolled. What is the expected value of this game?

$X(s) = 4 - s$ ← r.v. representing how much \$ you get

pmf:

x	-2	-1	0	1	2	3
$p_X(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$E[X] = \sum x\, p_X(x) = -2\left(\tfrac{1}{6}\right) - 1\left(\tfrac{1}{6}\right) + 0 + \tfrac{1}{6} + \tfrac{2}{6} + \tfrac{3}{6}$$

$$= \tfrac{1}{2}$$

Expected value of a function f of a discrete r.v. X

$$E[f(X)] = \sum_{x \text{ in range}(X)} p_X(x)\, f(X)$$

Properties: X, Y discrete r.v.'s, $c \in \mathbb{R}$:

$$E(c) = c \qquad E[cX] = E[X] \qquad E[X+Y] = E[X] + E[Y]$$

The variance of a discrete r.v. X, $\mathrm{Var}(X)$ or σ_X^2 is

$$\mathrm{Var}(X) = E\left[(X-\mu_X)^2\right] = \sum_{\text{range}(X)} (x-\mu_X)^2 p_X(x)$$

the standard deviation is $\sigma_X = \sqrt{\mathrm{Var}(X)}$

Note: $(x-\mu_X)^2$ is the squared distance between the outcome x and the mean (or expected value) μ_X, so the variance is the expected value of squared distance from the mean.
- a measure of spread.

Properties of variance

$\mathrm{Var}(X) \geq 0$

$\mathrm{Var}(aX+b) = a^2 \mathrm{Var}(X)$ $\quad a, b \in \mathbb{R}$

$\mathrm{Var}(X) = E[X^2] - (E[X])^2$ ← very useful formula.

Standard discrete r.v.s

A Bernoulli trial is any experiment with a binary sample space

$S = \{\text{success}, \text{failure}\}$

happy, sad etc.

Probabilities $P(\text{success}) = p$

$P(\text{failure}) = 1-p = q$

Define a random variable $Y: S \to \mathbb{R}$ by $Y(\text{success}) = 1$

$Y(\text{failure}) = 0$

Then Y has pmf $p_Y(x) = \begin{cases} 1-p & x=0 \\ p & x=1 \\ 0 & \text{else} \end{cases}$

If a random variable X has the above pmf for some p we say it has the Bernoulli distribution $X \sim \mathrm{Bern}(p)$

The Bernoulli distribution has $E[X] = p$ and $Var[X] = pq$

If X is the random variable representing the number of successes in n independent Bernoulli trials with success probability p then X is a Binomial random variable, $X \sim Bin(n, p)$ with pmf

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

(combinations: $\binom{n}{x} = {}^nC_x$)

Continuous random variables

A continuous random variable is a random variable whose range is an infinite, uncountable subset of $\mathbb{R}$.

(eg: $[0,1]$ – there are infinitely many numbers here and you can't count them)

Example

What is the probability that my height is exactly 186cm?

last time I measured it was ~186cm, so let's say:

$$P(185 < H < 187) = 1$$

Suppose I assign $P(H = 186) = \frac{1}{2}$

and $P(H = 186.1) = \frac{1}{4}$

Problem: there are infinitely many numbers between 186 and 186.1

eg: 186.05, 186.025, 186.001, 186.0001, ...

we can't assign all of these a non-zero probability

why not? total probability will exceed 1

(and it doesn't make sense to assign 186.1 a non-zero probability but not, say 186.025.)

Solution: work with probability densities instead.

A probability density function (pdf) is a function $f : \mathbb{R} \to \mathbb{R}$ such that

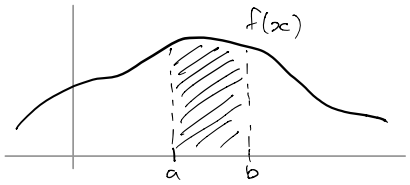
1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$ improper integral

If X is a continuous random variable such that

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

we say X has probability density function f, denote f_X

i.e: probabilities correspond to areas under f:



Note: $P(X = a) = 0$

$$P(a \le X \le b) = P(a < X \le b) = P(a < X < b)$$

FOR A CRV X WITH PDF f_X:

cumulative distribution function (cdf)

$F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(t) = P(X \le t) = \int_{-\infty}^{t} f_X(x)\,dx$$

expected value

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x\, f_X(x)\,dx$$

variance

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X-\mu_X)^2] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx$$

compare - discrete versions:

$$\sum_{x \in \text{range } X} x\, p_X(x)$$

$$\sum (x-\mu_X)^2 p_X(x)$$

Properties of $E[X]$ and $\mathrm{Var}[X]$ also hold for crv's.

$E(c) = c$ $\qquad E[cX] = E[X]$ $\qquad E[X+Y] = E[X] + E[Y]$

$\mathrm{Var}(X) \ge 0$

$\mathrm{Var}(aX+b) = a^2 \mathrm{Var}(X)$ $\qquad a, b \in \mathbb{R}$

$\mathrm{Var}(X) = E[X^2] - (E[X])^2$ ← very useful formula.

Some useful pdf's

The uniform distribution

Suppose X is a crv with range $X = [a, b] \subset \mathbb{R}$

define $$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

f is a pdf called the uniform distribution

check: $f(x) \ge 0$? $b > a$ ✓

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \left[\frac{x}{b-a}\right]_a^b = \frac{b}{b-a} - \frac{a}{b-a} = \frac{b-a}{b-a} = 1$$

Example: spinning a wheel

model probability of stopping with orientation within a given range.

$[0, 2\pi]$ $$f(x) = \begin{cases} \frac{1}{2\pi} & 0 \le x \le 2\pi \\ 0 & \text{otherwise} \end{cases}$$

for $a \in [0, 2\pi]$, $\varepsilon > 0$

$$P(a-\varepsilon < X < a+\varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} \frac{1}{2\pi}\,dx = \left[\frac{x}{2\pi}\right]_{a-\varepsilon}^{a+\varepsilon}$$

$$= \frac{a+\varepsilon}{2\pi} - \frac{(a-\varepsilon)}{2\pi}$$

$$= \frac{\varepsilon}{\pi}$$

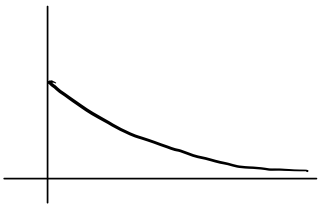
i.e. the probability of the wheel stopping in a given interval depends only on the size of the interval - not where it is hence "uniform" distribution.

The exponential distribution

Let T be a continuous random variable which takes positive values (range $T = [0, \infty)$)

define, for any $\lambda \geq 0$,

$$f_T(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$



then f_T is a pdf called the exponential distribution.

check: $f_T(x) \geq 0$? ✓ $\lambda \geq 0$, $e^a > 0$

$$\int_{-\infty}^{\infty} f_T(x) = \int_0^{\infty} \lambda e^{-\lambda x} dx = \lim_{t \to \infty} \int_0^t \lambda e^{-\lambda x} dx = \lim_{t \to \infty} \left[-e^{-\lambda x} \right]_0^t$$

$$= \lim_{t \to \infty} \left(-e^{-\lambda t} + 1 \right) = 0 + 1 = 1$$

FOR THE EXPONENTIAL DISTRIBUTION: $T \sim \exp(\lambda)$

cdf: $F_T(t) = 1 - e^{-\lambda t}$

expected value $\mu_T = \frac{1}{\lambda}$

variance $\mathrm{Var}(T) = \frac{1}{\lambda^2}$

$\sigma_T = \frac{1}{\lambda}$

The exponential distribution is used to model things like wait times.

$\mu_T = \frac{1}{\lambda}$ so as λ increases the expected value for wait time decreases → think of λ as the "rate" of occurrence

The normal distribution

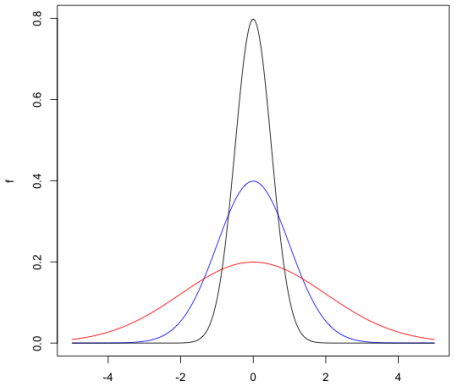
Let μ, σ be some parameters and define

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

then f_X is a pdf called the normal (or Gaussian) distribution (won't prove this)

If X has the normal distribution we write $X \sim N(\mu, \sigma^2)$ and then, as the notation suggests:

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$



$\mu = 0$ $\sigma = 0.5$

$\sigma = 1$

$\sigma = 2$

Used to model things like height in a random sample of people.

The standard normal distribution is $N(0,1)$ $(\mu = 1, \sigma = 1)$

If X is a random variable with $X \sim N(\mu, \sigma_x^2)$ then

$Z = \frac{X - \mu_x}{\sigma_x}$ is a r.v. with the standard normal dist.

i.e. $Z \sim N(0,1)$. Applying this transformation $Z = \frac{X-\mu_x}{\sigma_x}$ is called standardising the r.v. X.

Statistical inference

- using random samples to estimate model parameters

eg: We might model human heights using a normal distribution, so we need a mean μ and standard deviation σ.
Not practical to measure every human, so we can't get μ exactly, but we could estimate it by taking a group of 20 people and calculating the mean height of those 20 people.

- how good would this estimate be? If we took another random sample of 20 people how close would the new estimate be to the old one?

A random sample is a collection of random variables $X_1, X_2 \dots, X_n$ which are independent and identically distributed (iid) i.e. each X_i has the same probability distribution, which we will call the common distribution, and is independent of all the other X_j, $j \neq i$.

Examples

- roll a die 10 times, X_i = outcome of the i^{th} roll
- select 20 people at random, H_i = height of the i^{th} person.
- drug trial on 20 people, E_i = effect of the drug on i^{th} person

A function of the elements X_i of a random sample which does not depend on unknown parameters of the common distribution for X_i is called a statistic

eg: $X_1 + X_2$, $\sum_{i=1}^{n} X_i$, $2X_n$, $\sum \frac{X_i}{n}$ are all statistics

$X_1 + \mu_X$, $\sum (X_i - \mu_X)$, X_n / σ_X are not

An **estimator** is a statistic that can be used to provide an **estimate** of a model parameter.

i.e. estimator : is a rule/formula

(point) estimate : is the result of applying the rule to an observation of the random sample

An estimator for the mean μ_X is $\bar{X}_n = \dfrac{X_1 + X_2 + \dots + X_n}{n}$

note: this is the only estimator we will discuss in detail in this unit

Example H = height of a person

$H_1, H_2, \dots, H_n$ random sample

$h_1, h_2, \dots, h_n$ measurements of heights

μ_H mean height of humans model parameter

$\bar{H}_n$ estimator for mean human height statistic / est.

$\bar{h}_n = \dfrac{h_1 + h_2 + \dots h_n}{n}$ mean of measured heights point estimate

the point estimate will usually be different for different random samples

The distribution of $\bar{X}_n$

X_i i.i.d random variables.

The estimator $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_n$ is a function of random variables, and so $\bar{X}_n$ is itself a random variable!

It therefore has: a probability distribution
an expected value
a variance

Expected value

Suppose X_i, $i = 1, \dots, n$ are iid rv's with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$

Then by the properties of $E[X]$:

$$E[\bar{X}_n] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{1}{n}\left(\underbrace{E[X_1]}_{\mu} + \underbrace{E[X_2]}_{\mu} + \dots + \underbrace{E[X_n]}_{\mu}\right)$$

$$= \frac{1}{n}(n\mu)$$

$$= \mu$$

Variance

recalling properties of $Var[X]$, in particular $Var[aX] = a^2 Var[X]$

$$Var[\bar{X}_n] = Var\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{1}{n^2}\left(Var[X_1] + \dots + Var[X_n]\right)$$

$$= \frac{1}{n^2}(n\sigma^2)$$

$$= \frac{\sigma^2}{n}$$

Observe: as sample size (n) increases, the variance of the estimator $\bar{X}_n$ decreases. Large samples → tighter estimates

Probability distribution

- Theorem: If the common distribution is Normal: $X_i \sim N(\mu, \sigma^2)$ then $\bar{X}_n$ also has the normal distribution, but with parameters $\mu, \frac{\sigma^2}{n}$, i.e. $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

- Central limit theorem (loosely) If the common distribution is anything with mean μ variance σ^2, then for large n the distribution of $\bar{X}_n$ is approximately normal $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ and the approximation improves as n increases.

Confidence Intervals

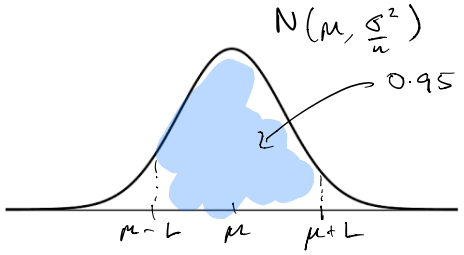
Let $X_1, X_2, \dots, X_n$ be i.id random variables such that the mean estimator $\bar{X}_n = \sum_i \frac{X_i}{n}$ is normally or approximately normally distributed:

$$\bar{X}_n \sim N(\mu, \tfrac{\sigma^2}{n}) \qquad (\mu \text{ unknown})$$

Suppose also that we know the variance σ^2 of the common distribution (or at least have an estimate we are convinced is good)

In terms of the unknown μ, what are the likely outcomes for $\bar{X}_n$? (i.e. likely point estimates of the mean)

Let's suppose that by "likely" we mean "within a range which has probability 0.95"



$$0.95 = P(\mu - L < \bar{X}_n < \mu + L)$$

standardise:

$$= P\left(\frac{-L}{\sigma/\sqrt{n}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{L}{\sigma/\sqrt{n}}\right)$$

$$0.95 = P\left(\frac{-L}{\sigma/\sqrt{n}} < Z < \frac{L}{\sigma/\sqrt{n}}\right) \qquad (Z \sim N(0,1))$$

The value of $\frac{L}{\sigma/\sqrt{n}}$ which satisfies $\uparrow$ is called the 0.975-quantile, denoted $z_{0.025}$, it is ≈ 1.96

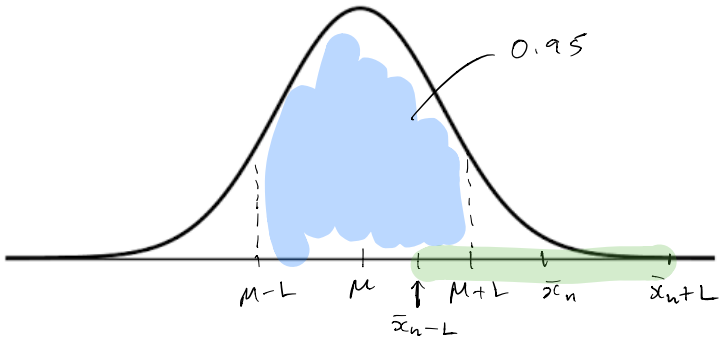
so $\frac{L}{\sigma/\sqrt{n}} = 1.96 \Rightarrow L = 1.96 \frac{\sigma}{\sqrt{n}}$

therefore the range of values with prob. 0.95 is $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$

Now if we have a point estimate $\bar{x}_n$ for $\bar{X}_n$ (i.e. we take a random sample and its mean = $\bar{x}_n$) then we define the 0.95 confidence interval for x_n as

$$\left(\bar{x}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x}_n + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

What does this μ?

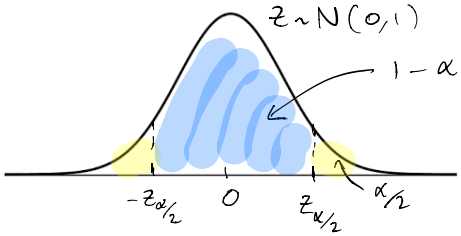


confidence interval

If the true value of the mean is outside the 95% C.I. for $\bar{x}_n$, then something unlikely has occurred: $\bar{x}_n$ is outside $(\mu - L, \mu + L)$ - an event which has probability less than 0.05

disclaimer:
- unlikely things sometimes happen.
- this is <u>not</u> the same as saying the probability that the true value of the mean is inside the C.I. is 0.95.

general $(1 - \frac{\alpha}{2})$-quantile:



general: $(1-\alpha)$ confidence interval:

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Hypothesis testing

Example: testing for loaded dice.

Fair die: $P(X=x) = \frac{1}{6} \quad x = 1, 2, \ldots, 6$

$$E[X] = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = 3.5$$

$$E[X^2] = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} + \ldots + \frac{36}{6} \simeq 15.16$$

$$\text{Var}[X] = E[X^2] - E[X]^2 \approx 15.16 - 3.5^2 = 2.9$$

Suppose you roll a die 20 times and observe a mean of 4.1 can you reasonably conclude that the die is loaded toward large numbers?

Null hypothesis $H_0: \mu = 3.5$ (the die is fair)

Alternative hypothesis $H_1: \mu > 3.5$ (high loaded die)

to test the hypothesis, assume H_0 is true. Then by CLT the dist. of $\bar{X}_n$ is approximately $N(\mu, \frac{\sigma^2}{n}) = N(3.5, \frac{2.9}{20})$

standardise $z = \dfrac{\bar{X}_n - 3.5}{\sqrt{2.9}/\sqrt{20}} \sim N(0,1)$ 'test statistic'

- under the null hypothesis, the expected value of z is 0.

the observed test statistic is

$$z_0 = \frac{4.1 - 3.5}{\sqrt{2.9}/\sqrt{20}} = 1.58$$

- this is saying that the point est 4.1 is 1.58 standard deviations above the hypothesized mean of 3.5.

now we need to figure out if this outcome is significantly GREATER than what we would expect under $H_0: \mu = 3.5$ (i.e. 0)

we choose a significance level $\alpha = 0.05$. This means we are defining significantly greater as being greater than $z_\alpha = 1.645$

(in this context, z_α is called the critical value)

(i.e we would deem the point estimate to be significantly greater than 3.5 if it were 1.645 standard deviations above.)

Since our observed test statistic 1.58 is less than the critical value 1.645, we decide that our evidence is consistent with the null hypothesis.

Notes

The above is called a one-sided test, because $H_1: \mu \geq 3.5$ we checked if the point estimate was significantly greater

To test the hypothesis $\mu \neq 3.5$ instead, we would check whether the point estimate is significantly different from 3.5, i.e. whether it is greater than $z_{\alpha/2}$ or less than $-z_{\alpha/2}$

(two sided test - the convention is to split α evenly between the two sides).