Lesson 16

Thursday 3/28/24

If x is a binomial variable, then $p(x)$ is:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

So, if p = 0.2 (probability of an event occurring in a Bernoulli trial) and we have n = 3 Bernoulli trials, then, $p(x)$ is:

$$p(x=0) = \frac{3!}{0!(3-0)!} p^0 (1-p)^{3-0} = \frac{6}{1 \times 6} \times 1 \times 0.8^3 = 0.512$$

(Continued)

So, if p = 0.2 (probability of an event occurring in a Bernoulli trial) and we have n = 3 Bernoulli trials, then, $p(x)$ is:

$$p(x=0) = \frac{3!}{0!(3-0)!}p^0(1-p)^{3-0} = \frac{6}{1\times 6} \times 1 \times 0.8^3 = 0.512$$

$$p(x=1) = \frac{3!}{1!(3-1)!}p^1(1-p)^{3-1} = \frac{6}{1\times 2} \times 0.2^1 \times 0.8^2 = 3 \times 0.128 = 0.384$$

$$p(x=2) = \frac{3!}{2!(3-2)!}p^2(1-p)^{3-2} = \frac{6}{2\times 1} \times 0.2^2 \times 0.8^1 = 3 \times 0.032 = 0.096$$

$$p(x=3) = \frac{3!}{3!(3-3)!}p^3(1-p)^{3-3} = \frac{6}{6\times 1} \times 0.2^3 \times 0.8^0 = 1 \times 0.008 = 0.008$$

(Continued)

So, if p = 0.2 (probability of an event occurring in a Bernoulli trial) and we have n = 3 Bernoulli trials, then, $p(x)$ is:

x	p(x)
0	0.512
1	0.384
2	0.096
3	0.008

$$\sum_{x=0}^{3} p(x) = 1$$

R code:

```
> n=3
> x=0:3
> p=0.2
> px = choose(n,x)*p^x*(1-p)^(n-x)
> data.frame(x,px)
  x    px
1 0 0.512
2 1 0.384
3 2 0.096
4 3 0.008
```

R Example

x	p(x)
0	0.512
1	0.384
2	0.096
3	0.008

$$\sum_{x=0}^{3} p(x) = 1$$

Hypothesis Testing

Page 169 of your book:

> "Recall that we begin our hypothesis test by assuming that <u>the null hypothesis is true</u>..."

> "What we do, then, is <u>maintain our belief in the null hypothesis</u> until we are informed by the data that this belief is improbable given what we have observed." [Emphasis added]

Step 1: State the hypothesis

Usually this means describing the value we expect a parameter to have if the hypothesis is true.

Step 2: Decide What Evidence Would Convince Us That the Hypothesis is Wrong

* Usually this means specifying a probability distribution for the evidence assuming the hypothesized parameter value is correct.

* Then we decide what data would convince us that the hypothesis is wrong in light of that probability distribution. The zone of data values that would cause us to conclude the hypothesis is wrong is called the <u>critical region</u>.

* Generally, we require <u>very strong</u> evidence to convince us that a hypothesis is wrong.

* Your book (p. 168) uses the analogy of a criminal trial where a jury starts from a position of innocence and requires strong evidence to abandon that position and conclude the person is guilty.

Step 3: Collect Data

* This is the focus of a good deal of research attention (what you will learn about in research methods).

* We define our key concepts (i.e., what does arrest mean, what does recidivism mean, what does strain mean, etc.)

* We carry out surveys or collect administrative records.

* We conduct appropriate measurements.

Step 4: Analyze Data

* In this step, we study the information (both central tendency and variation) in our data.

* We compare the information we have obtained to the probability distribution that we developed in step 2.

* A key feature of this step is the estimation of a p-value: *the probability the evidence lies in the critical region if the hypothesis we are testing is true*.

* A commonly used *p*-value in social science research is 0.05. Sometimes people call this a *statistical significance test*.

Step 5: Make a Decision

* This is the time when we draw a conclusion about the status of our hypothesis.

* If the *p*-value we estimated in Step 4 is small (a typical level is 0.05 or less), we reject the hypothesis that is being tested.

* We recognize that this decision is made with uncertainty (see the table at the bottom of page 169 in your book).

* We could make a Type 1 error (rejecting a hypothesis when the hypothesis is true) or a Type 2 error (failing to reject a hypothesis when the hypothesis is false).

Let's consider an example (Cook & Zarkin, 1985):

CRIME AND THE BUSINESS CYCLE

*PHILIP J. COOK and GARY A. ZARKIN**

THE business cycle has a pervasive effect on the structure of economic opportunity and hence on behavior. The effect is reflected in social indicators as diverse as school enrollments, birthrates, and labor force participation.[1] It would be surprising indeed if crime rates were immune to general business conditions, and certainly the conventional wisdom asserts that "street" crime is countercyclical. A recession always provides police chiefs with a comfortable explanation for their failure to prevent increases in the crime rate.

Question: when the economy contracts, what happens to crime?

From the 1930's to the early 1980's, there were 9 complete business cycles (NBER Business Cycle Dating Committee)

August 1929 (1929Q3)	March 1933 (1933Q1)	43	21	64	34
May 1937 (1937Q2)	June 1938 (1938Q2)	13	50	63	93
February 1945 (1945Q1)	October 1945 (1945Q4)	8	80	88	93
November 1948 (1948Q4)	October 1949 (1949Q4)	11	37	48	45
July 1953 (1953Q2)	May 1954 (1954Q2)	10	45	55	56
August 1957 (1957Q3)	April 1958 (1958Q2)	8	39	47	49
April 1960 (1960Q2)	February 1961 (1961Q1)	10	24	34	32
December 1969 (1969Q4)	November 1970 (1970Q4)	11	106	117	116
November 1973 (1973Q4)	March 1975 (1975Q1)	16	36	52	47
January 1980 (1980Q1)	July 1980 (1980Q3)	6	58	64	74
July 1981 (1981Q3)	November 1982 (1982Q4)	16	12	28	18

Cook and Zarkin Framework

- When each business cycle ends, the economy transitions from a growth phase to a contraction phase (a recession).
- For each of the 9 times the economy tipped into a recession, C&Z asked the question, "Did robbery rates move higher relative to their yearly trends during the growth phase?"
- Notice this is a yes or no question (like a Bernoulli trial).

* Step 1: State the hypothesis.

* If there is no connection at all between changes in the economy and changes in crime, what would we expect to see?

* We might expect to see that when the economy tips into recession, robberies would be equally likely to increase or decrease (relative to their trend in the growth phase). This is similar to the idea that a coin flip would be equally likely to come up heads or tails.

* So each business cycle could be viewed as a separate Bernoulli trial with a "yes" (robbery increases) or "no" (robbery decreases) outcome.

* If our assumption is correct, we would expect to see robbery increasing about half the time and decreasing about half the time.

* So, we hypothesis that our Bernoulli/binomial parameter is equal to 0.5 (same as the probability of a heads or tails when we flip a coin).

Step 2: What evidence would convince us that our hypothesis (probability that robbery increases when economy tips into a recession = 0.5) is wrong?

$$p(x) = \binom{9}{x} 0.5^x (1-0.5)^{9-x}, \text{ for } x = 0, 1, \ldots, 9$$

	k	p(x)	
CR	0	0.002	p(0 or 1 robbery increases) = 0.002 + 0.018 = 0.020
	1	0.018	
	2	0.070	
	3	0.164	
	4	0.246	p(2-7 robbery increases) = 0.070+0.164+0.246+0.246+0.164+0.070 = 0.96
	5	0.246	
	6	0.164	
	7	0.070	
CR	8	0.018	p(8 or 9 robbery increases) = 0.018 + 0.002 = 0.020
	9	0.002	

In words, if the probability of robbery increasing is really 0.5, it would be surprising to see 0, 1, 8, or 9 robbery increases (where surprising means less than a 0.05 probability). We say that the events 0, 1, 8, or 9 are all in the critical region. So, if we get one of these outcomes, we will reject the hypothesis that the probability of robbery increasing = 0.5.

* Step 3: C&Z collected their data by looking at changes in UCR crime statistics connected to each of the 9 business cycles between the 1930's and 1980's.

* Step 4: C&Z discovered that -- in 8 out of the 9 business cycles -- robbery increased relative to its growth phase trend when the economy tipped into recession. In other words, they found what they called a strong countercyclical pattern.

* Step 5: The question is whether the evidence is strong enough to reject the hypothesis of equally likely outcomes. In Step 2, we decided that 8 robbery increases would qualify as evidence that would be strong enough to reject the hypothesis that the probability of robbery increasing when the economy tips into a recession = 0.5.

Substantive interpretation:

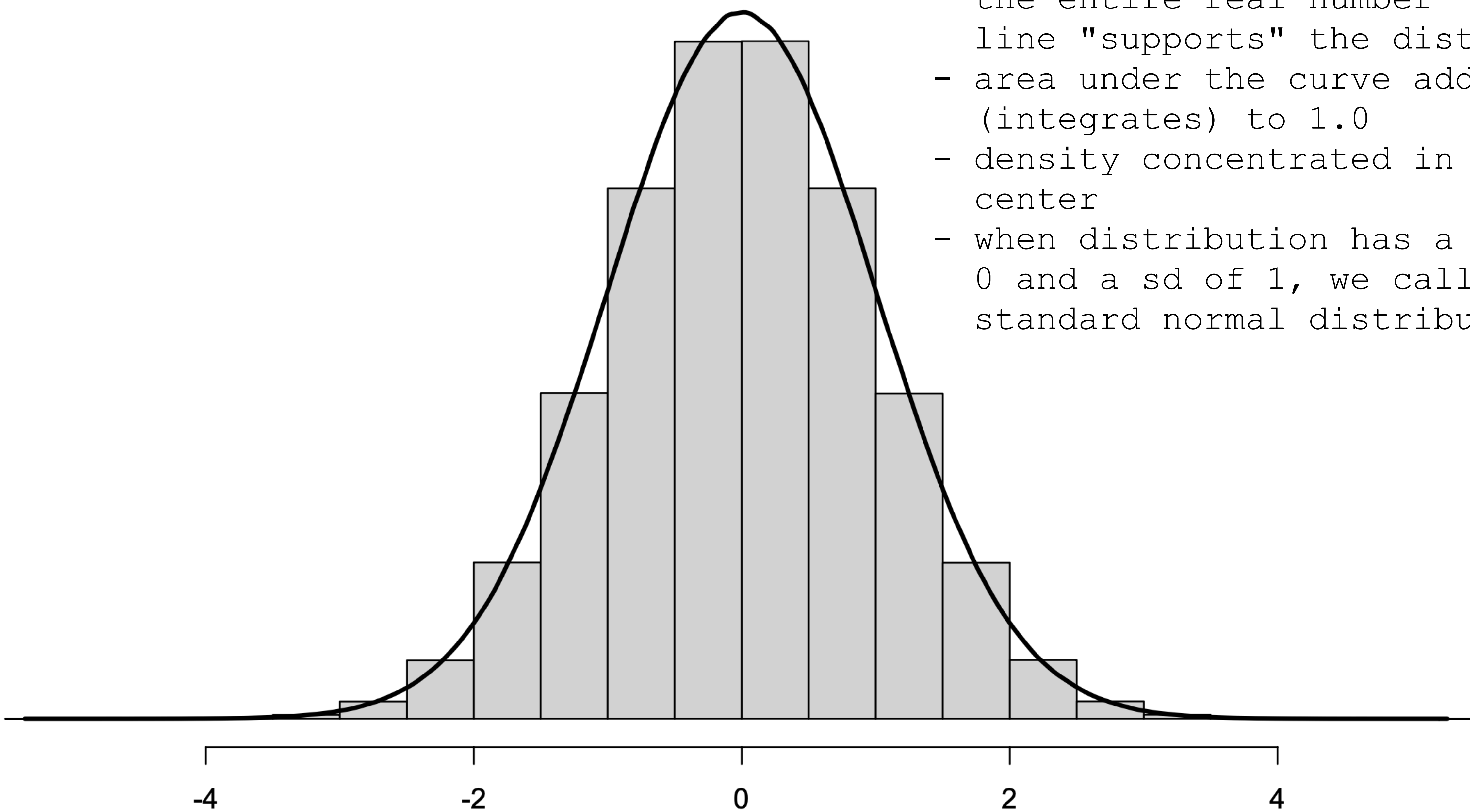
* Considering 9 business cycles, robbery increased when the economy tipped into a recession 8 times.

* One of two things has happened: (1) a rare event has occurred by chance (coincidence); or (2) the probability that robbery increases when the economy tips into a recession is not 0.5.

* Note that our inferential procedure is not strong enough to tell us which one has happened for sure. There will always be uncertainty.

* The probability that we could get a result in the *critical region*, x={0,1,8,9}, is a small number (0.04; if the hypothesis being tested is correct).

Normal Distributions



Key properties:

- continuous
- symmetric
- the entire real number line "supports" the distribution
- area under the curve adds up (integrates) to 1.0
- density concentrated in the center
- when distribution has a mean of 0 and a sd of 1, we call it the standard normal distribution

Normal Distributions (Continued)

- Because the distribution is continuous, it doesn't make sense to consider quantities like p(x = a specific number, $\boldsymbol{\tau}$); in fact $p(x = \boldsymbol{\tau}) = 0$.
- Instead, we can think about quantities like $p(x < \boldsymbol{\tau})$ or $p(x > \boldsymbol{\tau})$ or $p(\boldsymbol{\tau}_1 < x < \boldsymbol{\tau}_2)$
- Example: the standard normal distribution is symmetric with a mean of zero. So, $p(x > 0) = 0.5$ (and $p(x < 0)$ is also equal to 0.5). Note that $p(x < 0)+p(x > 0) = 1.0$.

Normal Probability Density Function

If x is a normally distributed variable, then the probability distribution (or density) of x is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

The parameters of this distribution are the mean (μ) and the standard deviation (σ).

Good news - you won't have to calculate this formula on exam!

Normal Distribution Properties

Suppose we have a long history of information about the waiting time to treatment for medical conditions at a corrections facility. The mean waiting time is 32 days, the standard deviation is 7 days and the waiting times are normally distributed.

- What is the estimated waiting time for someone who is 1 standard deviation above the mean?

Normal Distribution Properties

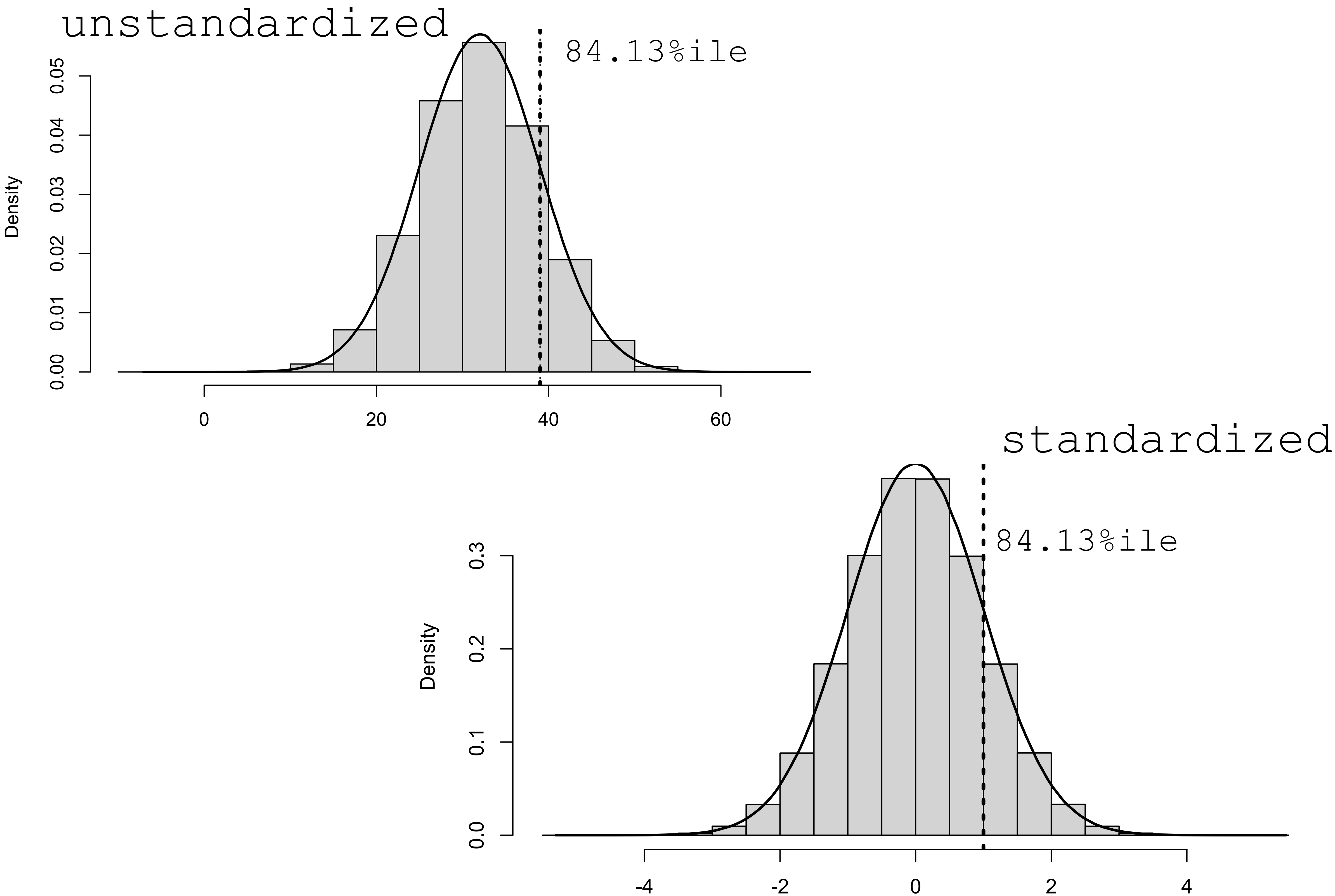
Considering our person, *i*, who is one standard deviation (7 days) above the mean (32 days), what is the z-score or standard score for that person?

$$z_i = \frac{x_i - \overline{X}}{s} = \frac{39 - 32}{7} = \frac{7}{7} = 1$$

Normal Distribution Properties

If we know person i is 1 standard deviation above the mean (a z-score of 1 means the person is 1 standard deviation above the mean) and we also know the waiting times are normally distributed, then we can get the percentile rank of that person's waiting time.

Step 1: calculate *z*-score (notice it is positive)
Step 2a: look at Table B.1 on page 533.
Step 2b: look at Figure 6.10 on page 176.
Step 3: find the entry for z=1.0
Step 4: notice that the table says 0.3413
Step 5: since z-score is positive, the percentile rank of person i is 0.5+0.3413 = 0.8413 or the 84.13%ile.


unstandardized
84.13%ile
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
standardized
84.13%ile
Density
0.0
0.1
0.2
0.3
-4
-2
0
2
4

Normal Distribution Properties (Example 2)

Suppose we have a long history of information about blood alcohol levels for people charged with DWI in a local jurisdiction. We know the data are normally distributed with a mean of 0.119 and a standard deviation of 0.008.

- What is the estimated BAC level for someone who is 1.25 standard deviations below the mean?

$0.119-(1.25*0.008) = 0.109$

Normal Distribution Properties

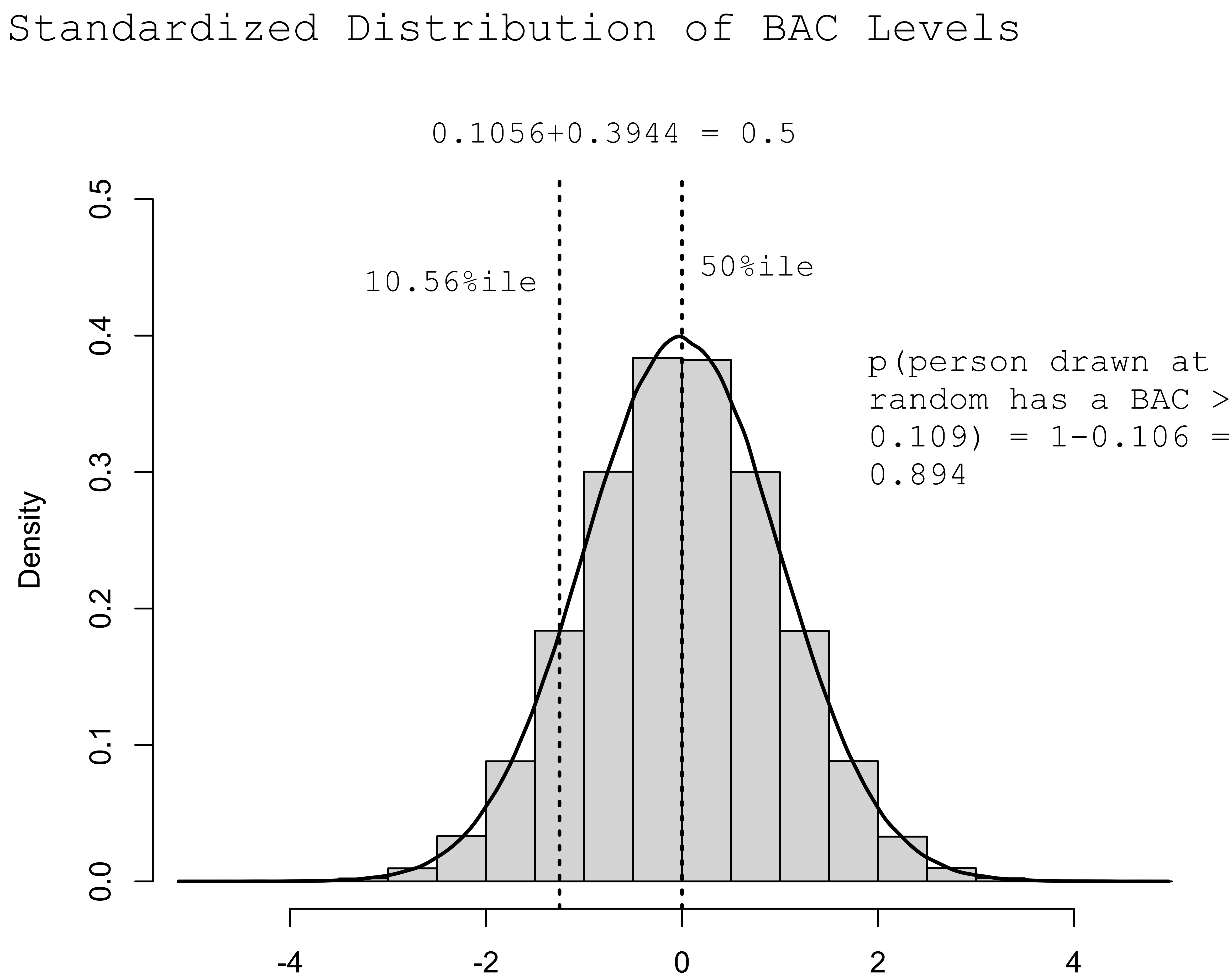
Considering our person, *i*, who has a BAC level of 0.109, what is the *z*-score or standard score for that person?

$$z_i = \frac{x_i - \overline{X}}{s} = \frac{0.109 - 0.119}{0.008} = \frac{-0.01}{0.008} = -1.25$$

Normal Distribution Properties

If we know person i is 1.25 standard deviations below the mean and we also know the BAC levels are normally distributed, then we can get the percentile rank of that person's BAC level.

Step 1: calculate z-score (notice it is negative)
Step 2: look at Table B.1 on page 533.
Step 3: find the entry for $z=-1.25$
Step 4: notice that the table says 0.3944
Step 5: since z-score is negative, the percentile rank of person i is $0.5-0.3944 = 0.1056$ or the 10.56%ile.


Standardized Distribution of BAC Levels
0.1056+0.3944 = 0.5
10.56%ile
50%ile
p(person drawn at random has a BAC > 0.109) = 1-0.106 = 0.894
Density
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4