Basic Statistics and Probability

Chapter 5: Continuous Random Variables

- Continuous Probability Distributions
- The Uniform Distribution
- The Normal Distribution
- Descriptive Methods for Assessing Normality
- Approximating the Binomial with a Normal

A continuous random variable is a random variable that can assume any value within some interval or intervals.

Examples:

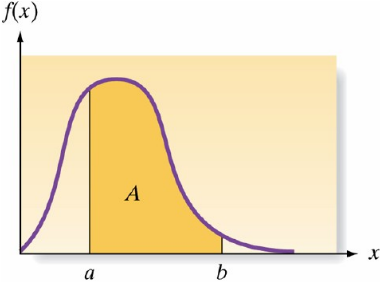
- The time a client has to wait in the bank until he/she is assisted.
- The length of a bird's leg.
- The blood glucose level of a person.
- Petrol consumption (in litres per 100 km) of a car.

Continuous Probability Distributions

The probability distribution of a continuous random variable X is characterized by a curve $f = f(x)$, called the probability density function.

Areas under the density are probabilities of events in terms of X:

Area A beneath f and between a and b
$=$ Probability that X is between a and b



For a continuous r.v. X, the probability that X is equal to a single value is 0:

$$P\{X = c\} = 0 \qquad \text{for all } c.$$

Consequently,

$$\text{Area } A = P\{a < X < b\} = P\{a \le X \le b\}.$$

Probability densities f have to fulfill:

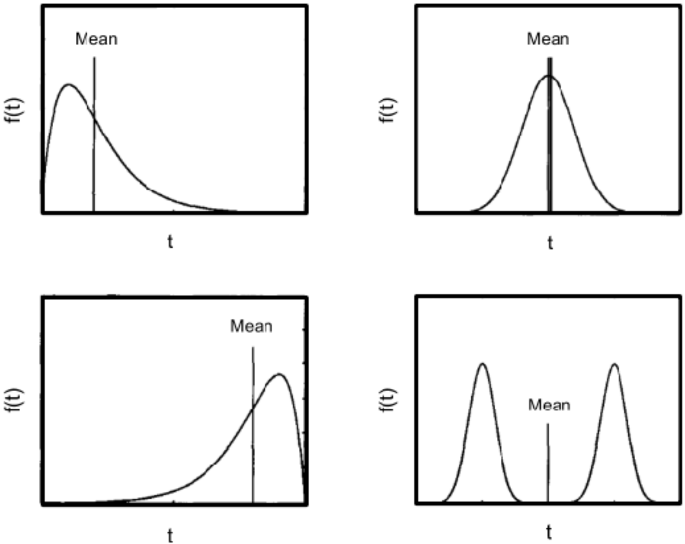
- $f(x) \ge 0$ for all x
- The total area under the density f should equal 1.

The areas of the most common continuous distributions are in tabular form or can be computed using statistical software like R.

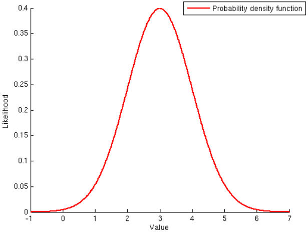
For Calculus experts.

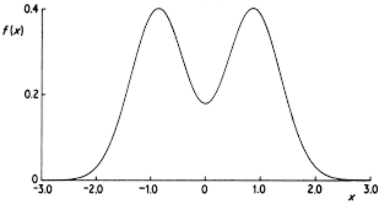
$$P(a < X < b) = \int_a^b f(x)dx$$

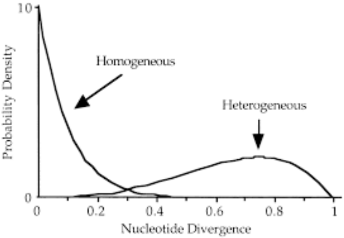
Probability densities can have different shapes, depending how likely are the values of the r.v.



Which are the most and least likely values of the r.v.'s with the following densities?



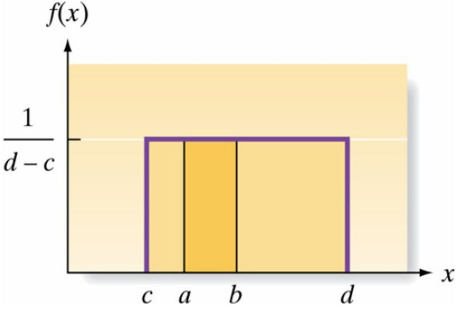




The Uniform Distribution

Random variables that appear to have equally likely outcomes over an interval follow a uniform probability distribution on that interval.

If X has a uniform distribution on an interval $[c, d]$, then its density f is constant and equal to $1/(d-c)$ over that interval, and f is 0 out of the interval.



In some sense: *all values in $[c, d]$ are equiprobable*

Example 5.1: Bus line

A bus line has a very regular frequency, stopping in my neighborhood every 10 minutes. If I arrive to the bus stop at random I will have to wait for less than 10 minutes. The waiting time X has a uniform distribution over the interval $[0, 10]$: $f(x) = \frac{1}{10}$ for $x \in [0, 10]$ ($f(x) = 0$ otherwise).

General rule for computing uniform probabilities:

If $X \sim \text{Unif}[c, d]$, then

$$P\{a < X < b\} = \frac{b-a}{d-c}, \quad \text{where } c \leq a < b \leq d.$$

Example 5.1 Bus line

Probability of waiting for more than 5 minutes:

$$P\{5 \leq X \leq 10\} =$$

$$\text{Expectation} = \mu = E(X) = \frac{c+d}{2}$$

$$\text{(Population) variance} = \sigma^2 = V(X) = \frac{(d-c)^2}{12}$$

$$\text{Standard deviation} = \sigma = \frac{d-c}{\sqrt{12}}$$

Example 5.1 Bus line

For Calculus experts:

$$\mu = E(X) = \int_c^d x\, f(x)\, dx$$

$$V(X) = E((X-\mu)^2) = E(X^2) - \mu^2 \quad \text{with} \quad E(X^2) = \int_c^d x^2 f(x)\, dx$$

Example 5.2: Uranium in Earth's crust

The trace amount (in ppm), X, of uranium in reservoirs follows a uniform distribution on the interval $[1, 3]$.

- Find $E[X]$. Interpret the value.

  $E[X] = \frac{3+1}{2} = 2$. The average trace amount is $X = 2$ ppm.

- Compute $P(2 < X < 2.5)$.

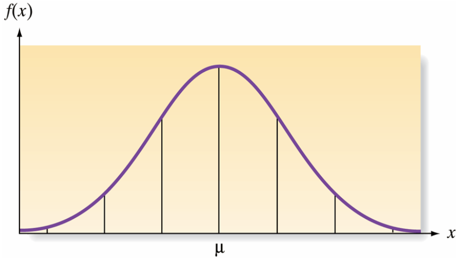
  $P(2 < X < 2.5) = (2.5 - 2) \cdot \frac{1}{2} = .25$.

- Compute $P(X < 1.75)$.

  $P(X < 1.75) = (1.75 - 1) \cdot \frac{1}{2} = .75 \times .5 = .375$.

The Normal Distribution

One of the most commonly observed continuous random variables has a bell-shaped probability density, called the bell curve:



It is the normal or Gaussian distribution, the most important continuous distribution.

It is a good approximation to the probability distribution of a random variable which results of adding many independent effects.

If we think about random biological or physical processes, they can often be viewed as being affected by a large number of random factors with individually small effects. The sum of all these random components produces an approximately normal random variable. For this reason, errors in measurements or some biometric data (such as height, weight, length, diameter,. . .) are usually assumed to be Gaussian.

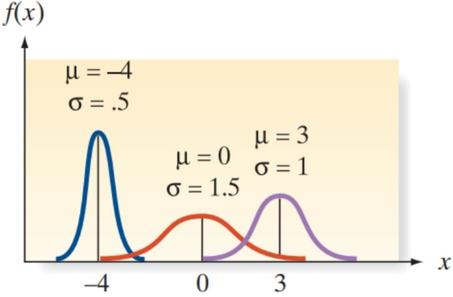
> “Suppose you bake 100 loaves of bread, each time following a recipe that is meant to produce a loaf weighing 1,000 grams. By chance you will sometimes add a bit more or a bit less flour or milk, or a bit more or less moisture may escape in the oven. If in the end each of a myriad of possible causes adds or subtracts a few grams, [...] the weight of your loaves will vary according to the normal distribution.” Mlodinow (2008)

A normally distributed r.v. can take any real value.

The density function of a normal r.v. X with expectation $E(X) = \mu$ and standard deviation $\sigma = \sqrt{V(X)}$ is

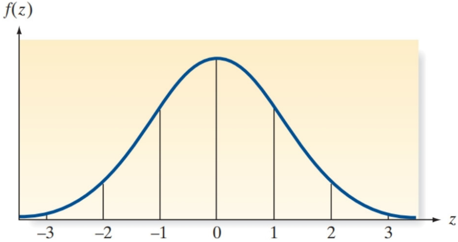
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

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



The normal density is symmetric about its mean μ. The higher σ is, the more spread the density is.

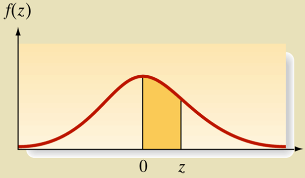
Computing the area over intervals under the normal probability distribution is a difficult task. Thus, we will use the computed areas listed in Table II of Appendix B, which refers to a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, called a standard normal distribution and denoted by $N(0, 1)$.



The probabilities of events involving a $N(\mu, \sigma)$ distribution will be expressed in terms of probabilities of a $N(0, 1)$. Then we will use Table II.

Table II in Appendix B

The 4-decimal numbers give the area (probability) between 0 and z.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

Examples of use of Table II: Let Z be a $N(0,1)$ r.v.

- $P\{0 < Z < 1\} = .3413$
- $P\{-1 < Z < 0\} = .3413$
- $$\begin{aligned} P\{.31 < Z < 1.15\} &= P\{0 < X < 1.15\} - P\{0 < X < .31\} \\ &= .3749 - .1217 = .2532 \end{aligned}$$
- $$\begin{aligned} &P\{-1.33 < Z < 1.33\} \\ &\qquad = P\{-1.33 < Z < 0\} + P\{0 < Z < 1.33\} \\ &\qquad = 2\,P\{0 < Z < 1.33\} = 2\,.4082 = 0.8164 \end{aligned}$$
- $P\{Z > 1.25\} = 0.5 - P\{0 < Z < 1.25\} = 0.5 - 0.3944 = 0.1056$
- $P\{Z < 0.58\} = 0.5 + P\{0 < Z < 0.58\} = 0.5 + 0.2190 = 0.719$
- $$\begin{aligned} P\{|Z| > 1.33\} &= P\{Z > 1.33\} + P\{Z < -1.33\} \\ &= P(\{-1.33 < Z < 1.33\}^c) \\ &= 1 - P\{-1.33 < Z < 1.33\} \\ &= 1 - 0.8164 = 0.1836 \end{aligned}$$

Converting $N(\mu,\sigma)$ probabilities to $N(0,1)$ ones

We will use the following property of the normal distribution:

$$\text{If} \quad X \sim N(\mu,\sigma) \quad \text{then} \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

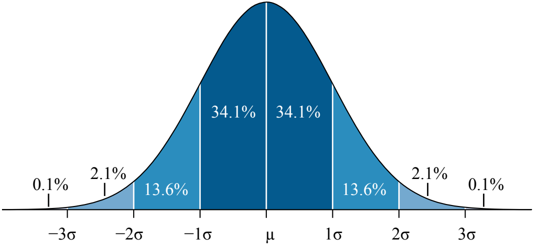
Example 5.3: For $X \sim N(2, 1.5)$ find $P\{-1 < X < 2.5\}$:

First observe that $Z = \dfrac{X-2}{1.5} \sim N(0,1)$. Then

$$\begin{aligned} P(-1 < X < 2.5) &= P\left(\frac{-1-2}{1.5} < Z < \frac{2.5-2}{1.5}\right) \\ &= P(-2 < Z < .33) = P(-2 < Z < 0) + P(0 < Z < .33) \\ &= P(0 < Z < 2) + P(0 < Z < .33) \\ &= .4772 + .1293 = .6065 \end{aligned}$$

How “the normal rule” is derived:

- $P(\mu - \sigma < X < \mu + \sigma) = P(-1 < Z < 1) = 2 \cdot (.3413) = .6826$
- $P(\mu{-}2\sigma < X < \mu{+}2\sigma) = P(-2 < Z < 2) = 2{\cdot}(.4772) = .9544$
- $P(\mu{-}3\sigma < X < \mu{+}3\sigma) = P(-3 < Z < 3) = 2{\cdot}(.4987) = .9974$



McClave's Example 5.8

A car manufacturer reports that one of its models has an in-city mileage (in miles per gallon) that follows a $N(27, 3)$ distribution. You have bought a car of this model and found that it only averages 20 miles per gallon. How unlikely is to get a car like this under the given assumptions?

$$\begin{aligned} P\{X < 20\} \quad &= P\left\{Z < \frac{20-27}{3}\right\} = P\{Z < -2.33\} \\ &= .5 - P(0 < Z < 2.33) = .5 - .4901 = .0099 \end{aligned}$$

This means that the probability of getting such a "lemon" is less than 1%.

Sometimes we are given a certain probability and we wish to find the values of the normal variable corresponding to that probability.

Example 5.4: For a $Z \sim N(0,1)$, find

- the 90th percentile, that is, the value $z_{0.1}$ leaving 0.9 probability to the left and 0.1 to the right

$$0.9 = P\{Z < z_{0.1}\} = 0.5 + P\{0 < Z < z_{0.1}\}$$

$$\Rightarrow 0.4 = P\{0 < Z < z_{0.1}\} \Rightarrow z_{0.1} = 1.28$$

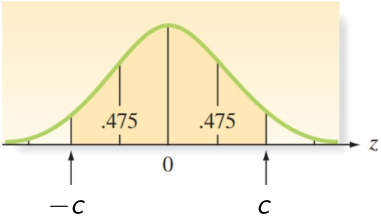
- the 95th percentile, that is, the value $z_{0.05}$ leaving 0.95 probability to the left and 0.05 to the right

$$0.95 = P\{Z < z_{0.05}\} = 0.5 + P\{0 < Z < z_{0.05}\}$$

$$\Rightarrow 0.45 = P\{0 < Z < z_{0.05}\} \Rightarrow z_{0.05} = 1.645$$

Example 5.4: Find the value c such that 95% of the standard normal Z values lie between $-c$ and c, that is, such that

$$0.95 = P\{-c < Z < c\}$$



$$\Rightarrow 0.475 = P\{0 < Z < c\}$$

$$\Rightarrow c = z_{0.025} = 1.96$$

Example 5.11 in McClave & Sincich: Suppose the scores X on a college entrance examination are normally distributed with a mean of 550 and a standard deviation of 100. A certain prestigious university will consider for admission only those applicants whose scores exceed the 90th percentile of the distribution. Find the minimum score an applicant must achieve in order to receive consideration for admission to the university.

Assessing Normality

Most of the Inferential Statistics methods that we will use require that the data under analysis come from a normal distribution.

If the data are clearly nonnormal, inferences derived from the method may be invalid.

Consequently, it is important to determine whether the sample comes from a Gaussian population before we can apply these techniques properly.

In general, we will not be able to affirm that they proceed from a normally distributed variable, but we will be able to discard data that are clearly "not normal".

We will introduce four descriptive methods which can be used to check for normality.

1. Construct a histogram or a stem-and-leaf display for the data. Note the shape of the graph. If the data are approximately normal, the graph will be bell-shaped (i.e., mound shaped and symmetric about the mean).
2. Compute the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$, and determine the percentage of observations falling into each. If the data are approximately normal, the percentages will be approximately equal to 68%, 95%, and 100%, respectively.
3. Find the interquartile range IQR and standard deviation s for the sample. Calculate the ratio IQR$/s$. If the data are approximately normal, then IQR$/s \simeq 1.3$.
4. Construct a normal probability plot for the data. If the data are approximately normal, the points will fall (approximately) on a straight line.

What is a normal probability plot?

It is a scatterplot with the sample observations ordered from smallest to largest on one axis and their corresponding quantiles from a standard normal distribution on the other axis.

We always use statistical softwares to generate a normal probability plot.

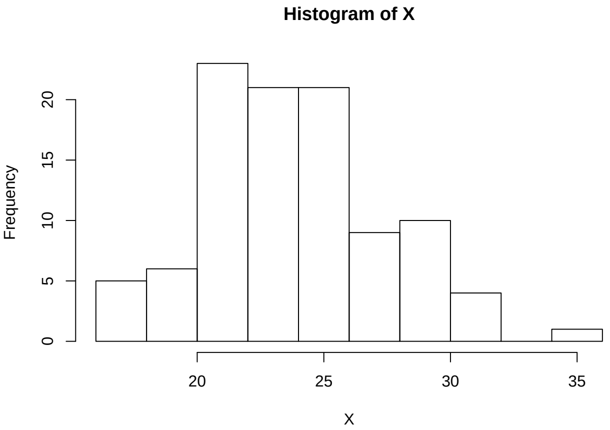
Example 5.5: Car speeds (Source: DASL)

John Beale (of Stanford) recorded the speeds of cars driving past his house, where the speed limit was 20 mph. He recorded every car for a two-month period. These are 100 representative readings.

16.27	16.57	16.70	17.17	17.84	18.50	18.59	18.71	18.88	19.11
19.64	20.10	20.20	20.28	20.62	20.64	20.65	20.73	20.76	20.82
21.17	21.22	21.23	21.53	21.54	21.57	21.60	21.75	21.78	21.79
21.89	21.91	21.97	21.97	22.33	22.35	22.41	22.47	22.58	22.73
22.96	23.04	23.07	23.11	23.25	23.43	23.45	23.46	23.50	23.52
23.53	23.54	23.59	23.75	23.86	24.02	24.07	24.12	24.18	24.19
24.23	24.26	24.30	24.56	24.76	25.19	25.21	25.30	25.45	25.49
25.52	25.60	25.70	25.70	25.71	25.81	26.07	26.09	26.33	26.65
27.28	27.34	27.49	27.61	27.87	28.06	28.28	28.45	28.81	28.96
28.98	29.51	29.53	29.94	29.97	30.03	30.10	30.72	31.26	34.06

Example 5.5 Car speeds

```
X = scan("car-speeds-100.txt")
hist(X)
```



Example 5.5 Car speeds

```
mean(X)
[1] 23.8439
sd(X)
[1] 3.563338
```

$$\bar{x} \pm s = 23.8439 \pm 3.5633 = (20.2806, 27.4072)$$

Relative frequency of $\bar{x} \pm s = \frac{68}{100} = 0.68$

$$\bar{x} \pm 2s = 23.8439 \pm 2 \cdot 3.5633 = (16.7172, 30.9706)$$

Relative frequency of $\bar{x} \pm 2s = \frac{95}{100} = 0.95$

$$\bar{x} \pm 3s = 23.8439 \pm 3 \cdot 3.5633 = (13.1539, 34.5339)$$

Relative frequency of $\bar{x} \pm 3s = \frac{100}{100} = 1$

Example 5.5 Car speeds

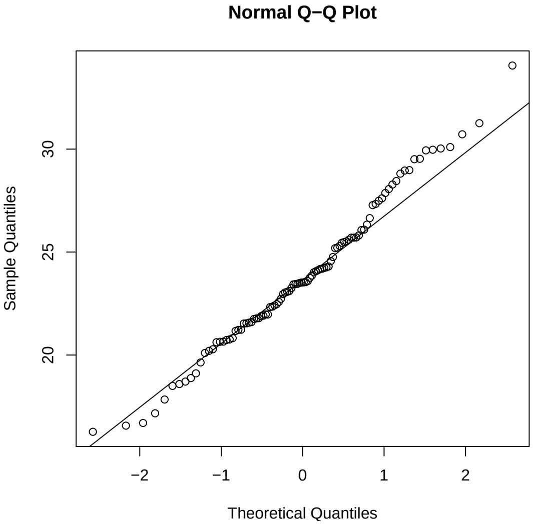
```
summary(X)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
16.27   21.56   23.52   23.84   25.73   34.06 
```

$$\text{IQR} = 25.73\text{-}21.56 = 4.17$$

$$\frac{\text{IQR}}{s} = \frac{4.17}{3.5633} = 1.17$$

Example 5.5 Car speeds

```
qqnorm(X)
qqline(X)
```



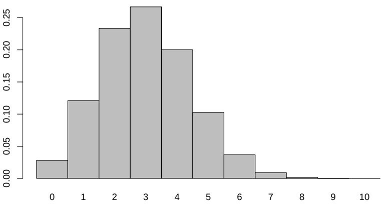
Approximating the Binomial with a Normal

When n is large, a normal probability distribution may be used to provide a good approximation to the probability distribution of a binomial random variable.

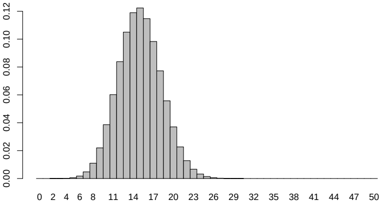
When n grows, bar diagrams for a $B(n, p)$ resemble the density function of a normal.

Example 5.6: For $p = 0.3$, the histogram for the mass function of a $B(n, p)$ is

$n = 5$



$n = 50$



This fact can be used to do calculations for binomials.

This is why binomial tables are only given for “small” values of n. For $X \sim B(n,p)$ and n “large enough” we have that

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx.}}{\sim} N(0,1)$$

As a rule of thumb, may use this normal approximation when $n \geq 20$, $p \geq 0.1$ and $1 - p \geq 0.1$.

Example in two attempts

Let $X \sim B(30, .3)$. Find $P(X \leq 7)$. Find $P(X \geq 8)$.

Observe: $\mu = 9$, $\sigma = \sqrt{30 \cdot (.3) \cdot (.7)} = 2.51$

First attempt, write

$$P(X \leq 7) = P\left(\frac{X-9}{2.51} \leq \frac{7-9}{2.51}\right) \approx P(Z \leq -.80) = .5 - .2881 = .2119$$

$$P(X \geq 8) = P\left(\frac{X-9}{2.51} \geq \frac{8-9}{2.51}\right) \approx P(Z \geq -.40) = .5 + .1554 = .6554$$

But $.6554 + .2119 = .8673 \neq 1$.

Reason: we are disregarding
$P(-.80 < Z < . - 40) = .2881 - .1554 = .1327$.

Example in two attempts Second attempt.

To avoid this situation we use what is called correction for continuity:

$$P(X \leq 7) = P(X \leq 7.5) = P\left(\frac{X-9}{2.51} \leq \frac{7.5-9}{2.51}\right)$$

$$\approx P(Z \leq -.60) = .5 - .2257 = .2743$$

$$P(X \geq 8) = P(X \geq 7.5) = P\left(\frac{X-9}{2.51} \geq \frac{7.5-9}{2.51}\right)$$

$$\approx P(Z \geq -.60) = .5 + .2257 = .7257$$