

# STA258H5

## University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

# THE CENTRAL LIMIT THEOREM

Important for Inference (later)

## Theorem 7.5

$$M_n(t) = E(e^{tx}) \quad \text{mgf}$$

Let  $Y$  and  $Y_1, Y_2, Y_3, \dots$  be random variables with moment-generating functions  $M(t)$  and  $M_1(t), M_2(t), \dots$ , respectively. If

$$\lim_{n \rightarrow \infty} M_n(t) = M(t) \quad \text{for all real } t,$$

then the distribution function of  $Y_n$  converges to the distribution function of  $Y$  as  $n \rightarrow \infty$ .

STA 256 convergence in distribution in terms  
of  $F_x(x)$   
CDF

$$\lim_{n \rightarrow \infty} F_x(x) = F(x)$$

↖ seq      ↖ converging RV

# Maclaurin Series

Taylor Series around  $x_0$

$$f(x) \approx f(x_0) + \frac{f'(x_0)}{1!} (x - x_0) + \frac{f''(x_0)}{2!} (x - x_0)^2 + \dots$$



A Maclaurin series is a Taylor series expansion of a function about  $0$ ,  $x_0 = 0$

$$f(x) = f(0) + f'(0)x + \frac{f''(0)x^2}{2!} + \frac{f^{(3)}(0)x^3}{3!} + \dots + \frac{f^{(n)}(0)x^n}{n!} + \dots$$

Maclaurin series are named after the Scottish mathematician Colin Maclaurin.

# Useful properties of MGFs

- $M_Y(0) = E(e^{0Y}) = E(1) = 1.$
- $M'_Y(0) = E(Y).$
- $M''_Y(0) = E(Y^2).$
- $M_{aY}(t) = E(e^{t(aY)}) = E(e^{(at)Y}) = M_Y(at),$  where  $a$  is a constant.

# Central Limit Theorem

Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed random variables with  $E(X_i) = \mu$  and  $V(X_i) = \sigma^2 < \infty$ . Define

$$U_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

Then the distribution function of  $U_n$  converges to the standard Normal distribution function as  $n \rightarrow \infty$ . That is,

$$\lim_{n \rightarrow \infty} P(U_n \leq u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ for all } u.$$

Let  $Z_i = \frac{X_i - \mu}{\sigma}$ . Note that  $E(Z_i) = 0$  and  $V(Z_i) = 1$ . Let us rewrite  $U_n$

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{\sigma} \right) = \sqrt{n} \left( \frac{\sum_{i=1}^n X_i - n\mu}{n\sigma} \right) = \frac{1}{\sqrt{n}} \left( \frac{\sum_{i=1}^n X_i - n\mu}{\sigma} \right)$$

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i.$$

Since the mgf of the sum of independent random variables is the product of their individual mgfs, if  $M_{Z_i}(t)$  denotes the mgf of each random variable  $Z_i$

$$M_{\sum_{i=1}^n Z_i}(t) = [M_{Z_1}(t)]^n$$

and

$$M_{U_n} = M_{\sum_{i=1}^n Z_i}(t/\sqrt{n}) = [M_{Z_1}(t/\sqrt{n})]^n.$$

Recall that  $M_{Z_i}(0) = 1$ ,  $M'_{Z_i}(0) = E(Z_i) = 0$ , and  $M''_{Z_i}(0) = E(Z_i^2) = V(Z_i^2) = 1$ .



Now, let us write the Taylor's series of  $M_{Z_i}(t)$  at 0

$$M_{Z_i}(t) = M_{Z_i}(0) + tM'_{Z_i}(0) + \frac{t^2}{2!}M''_{Z_i}(0) + \frac{t^3}{3!}M'''_{Z_i}(0) + \frac{t^4}{4!}M''''_{Z_i}(0) + \dots$$

$$M_{Z_i}(t) = 1 + \frac{t^2}{2} + \frac{t^3}{3!}M'''_{Z_i}(0) + \frac{t^4}{4!}M''''_{Z_i}(0) + \dots$$

$$M_{U_n}(t) = [M_{Z_1}(t/\sqrt{n})]^n = \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}M'''_{Z_i}(0) + \frac{t^4}{4!n^2}M''''_{Z_i}(0) + \dots\right]^n$$

Recall that if

$$\lim_{n \rightarrow \infty} b_n = b \qquad \lim_{n \rightarrow \infty} \left(1 + \frac{b_n}{n}\right)^n = e^b$$

But

$$\lim_{n \rightarrow \infty} \left[ \frac{t^2}{2} + \frac{t^3}{3!n^{1/2}} M_{Z_i}'''(0) + \frac{t^4}{4!n} M_{Z_i}''''(0) + \dots \right] = \frac{t^2}{2}$$

Therefore,

$$\lim_{n \rightarrow \infty} M_{U_n}(t) = \exp\left(\frac{t^2}{2}\right)$$

which is the moment-generating function for a standard Normal random variable. Applying Theorem 7.5 we conclude that  $U_n$  has a distribution function that converges to the distribution function of the standard Normal random variable.

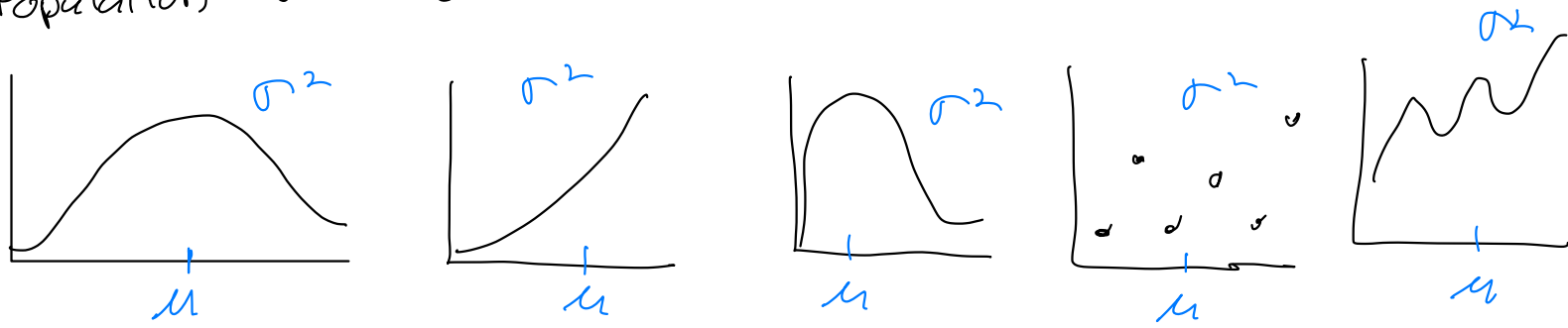
# Using the CLT

Two approximate distributions:

- 1  $\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$
- 2  $T = \sum_{i=1}^n X_i \approx N(n\mu, n\sigma^2)$

# Central Limit Theorem

Population  $\rightarrow$  follows a distribution (with a mean and var)



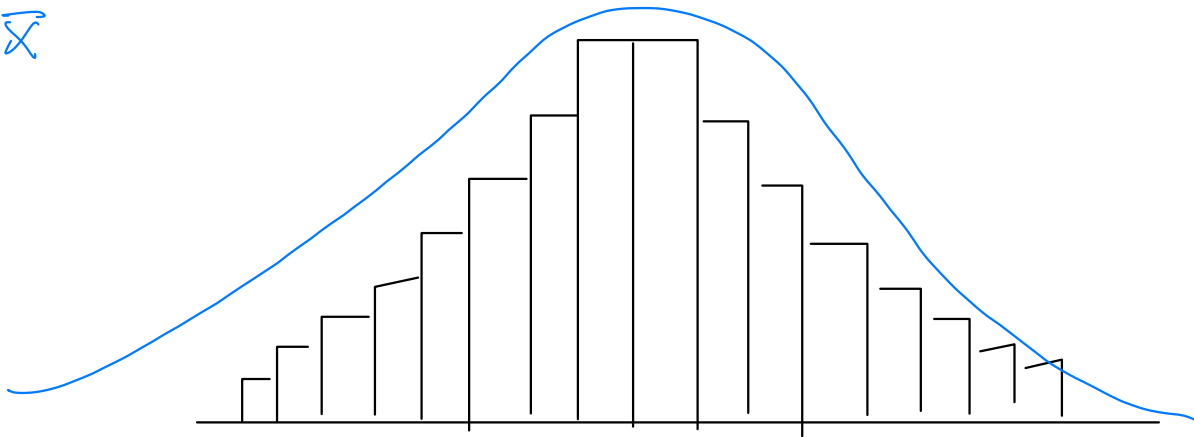
Take Samples  
of size  $n$   
from same pop.

...



Sampling Distribution  
of  $\bar{X}$

make histogram using  
the large collection of  $\bar{X}$ 's



$$\text{Mean: } \mu_{\bar{X}} = \mu$$

$$\text{Var: } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n})$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \begin{array}{l} \text{total} \\ \text{sample size} \end{array}$$

$$\underbrace{\sum_{i=1}^n x_i}_{\text{total}} = n\bar{x}$$

Sample mean

$$\bar{x} \sim N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n})$$

Consequence

$$n\bar{x} = \underbrace{\sum_{i=1}^n x_i}_T \sim N(\text{mean} = n\mu, \text{var} = n\sigma^2)$$

$$\text{var}(ax) = a^2 \text{var}(x)$$

$$\text{var}(n\sigma_{\bar{x}}^2) = n^2 \text{var}(\sigma_{\bar{x}}^2) = n \cdot \sigma^2$$

$$\bar{x}: Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leftarrow \text{st. dev}$$

$$\text{Total: } \underbrace{\sum x_i}_T \quad Z = \frac{T - n\mu}{\sqrt{n\sigma^2}} = \frac{T - n\mu}{(\sqrt{n})\sigma}$$

## Example

$$\mu = 2.2 \quad \sigma = 1.4 \quad n = 52$$
$$\sigma^2 = 1.96$$

The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution takes only whole-number values, so it is certainly not Normal.

- a) Let  $\bar{x}$  be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of  $\bar{x}$  according to the Central Limit Theorem?
- b) What is the approximate probability that  $\bar{x}$  is less than 2?
- c) What is the approximate probability that there are fewer than 100 accidents at the intersection in a year?

Example (slide 12)

a)  $\mu = 2.2$   $\sigma = 1.4$   $n = 52$   
 $\sigma^2 = 1.96$

want distrib of  $\bar{x}$

By CLT

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n})$$

$$\sim N(\mu_{\bar{X}} = 2.2, \sigma_{\bar{X}}^2 = \frac{1.96}{52})$$

$$\approx 0.0377$$

b)  $P(\bar{X} < 2)$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{2 - \mu}{\sigma/\sqrt{n}}\right)$$

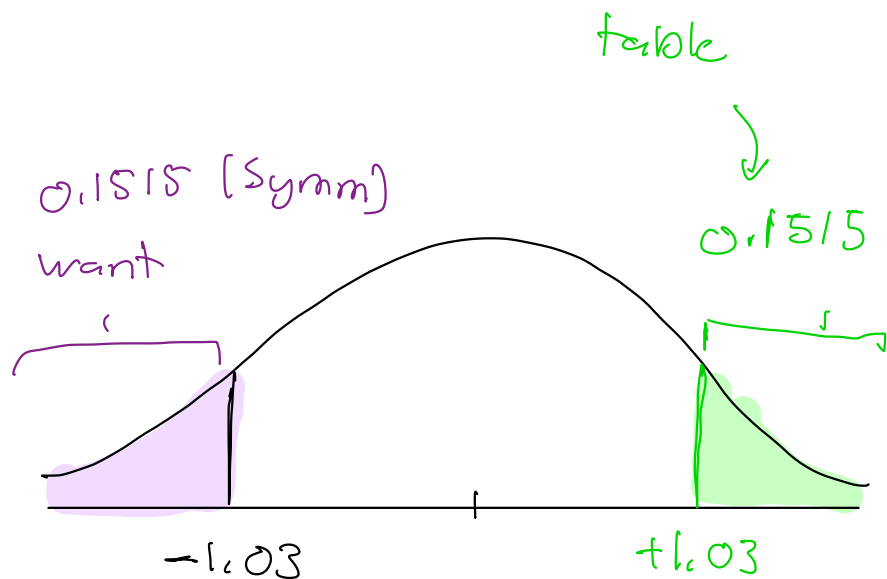
$\underbrace{\hspace{1cm}}_{=Z}$

$$\begin{cases} \mu = 2.2 \\ \sigma = 1.4 \\ n = 52 \end{cases}$$

$$= P\left(Z < \frac{2 - 2.2}{1.4/\sqrt{52}}\right)$$

$$= P(Z < -1.03)$$

$$= 0.1515$$



c) Fewer than 100 for the total of 52 weeks

$$\mu = 2.2 \quad \sigma = 1.4$$

$n$

$$T = \sum x_i \sim N(\text{mean} = n\mu, \text{var} = n\sigma^2)$$

$$T = \sum x_i \sim N(\text{mean} = (52)(2.2), \text{var} = (52)(1.4)^2)$$

$$\sim N(\text{mean} = 114.4, \text{var} = 101.92)$$

$$\text{var } n\sigma^2 = 101.92$$

$$\text{SD } \sqrt{n}\sigma = \sqrt{101.92}$$

$$T = \sum x_i$$

$$P(\text{Total} < 100)$$

$$\approx P(T < 100)$$

Totals

$$Z = \frac{T - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

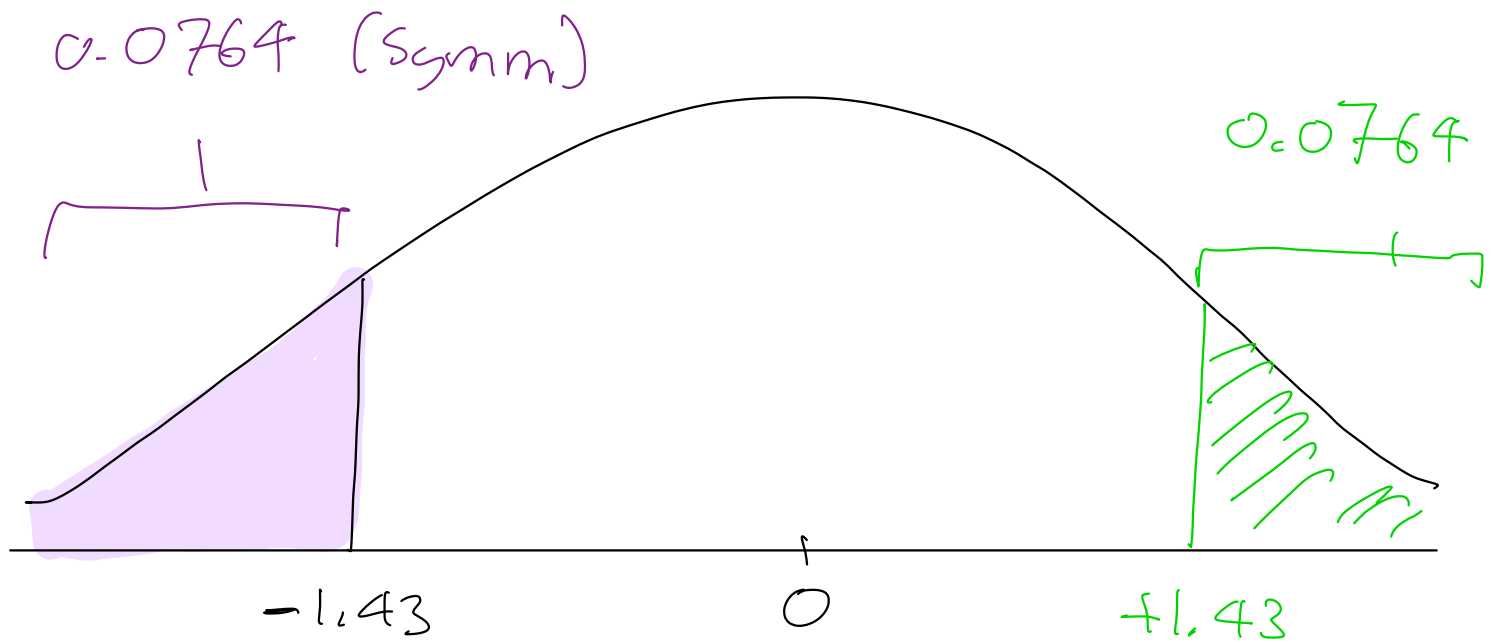
$$\approx P\left(\underbrace{\frac{T - n\mu}{\sqrt{n}\sigma}}_Z < \frac{100 - n\mu}{\sqrt{n}\sigma}\right)$$

$$= P\left(Z < \frac{100 - (52)(2.2)}{(\sqrt{52})(1.4)}\right)$$

$$\approx P(Z < -1.43)$$

$$= 0.0764$$





For  $\bar{X}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Totals:

$$T = \frac{(\bar{X} - \mu)n}{(\sigma/\sqrt{n})n} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

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


# Solution

a) By the Central Limit Theorem,  $\bar{X}$  is roughly Normal with mean  $\mu^* = 2.2$  and standard deviation  $\sigma^* = \sigma/\sqrt{n} = 1.4/\sqrt{52} = 0.1941$ .

$$\begin{aligned} \text{b) } P(\bar{X} < 2) &= P\left(\frac{\bar{X} - \mu^*}{\sigma^*} < \frac{2 - 2.2}{0.1941}\right) \\ &\approx P(Z < -1.0303) = 0.1515 \end{aligned}$$

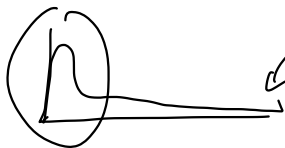
# Solution

Let  $X_i$  be the number of accidents during week  $i$ .

$$\begin{aligned} \text{c) } P(\text{Total} < 100) &= P\left(\sum_{i=1}^{52} X_i < 100\right) \\ &= P\left(\frac{\sum_{i=1}^{52} X_i}{52} < \frac{100}{52}\right) \\ &= P(\bar{X} < 1.9230) \\ &\approx P(Z < -1.4270) = 0.0768 \end{aligned}$$


## More on insurance

many have  
no claim or  
small claim



few have  
large claim

An insurance company knows that in the entire population of millions of apartment owners, the mean annual loss from damage is  $\mu = 75$  and the standard deviation of the loss is  $\sigma = 300$ . The distribution of losses is strongly right-skewed: most policies have \$0 loss, but a few have large losses. If the company sells 10,000 policies, can it safely base its rates on the assumption that its average loss will be no greater than \$85?

Example Slide 15)

$$\mu = 75$$

$$SD: \sigma = 300$$

$$n = 10,000$$

$$Var: \sigma^2 = 300^2 = 90000$$

Determine prob  $\underbrace{\text{avg loss}}_{\bar{x}}$   $\underbrace{\text{no greater than}}_{\leq}$  \$85

By CLT

$$\bar{x} \sim N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n})$$

$$\sim N(\mu_{\bar{x}} = 75, \sigma_{\bar{x}}^2 = \frac{300^2}{10000})$$

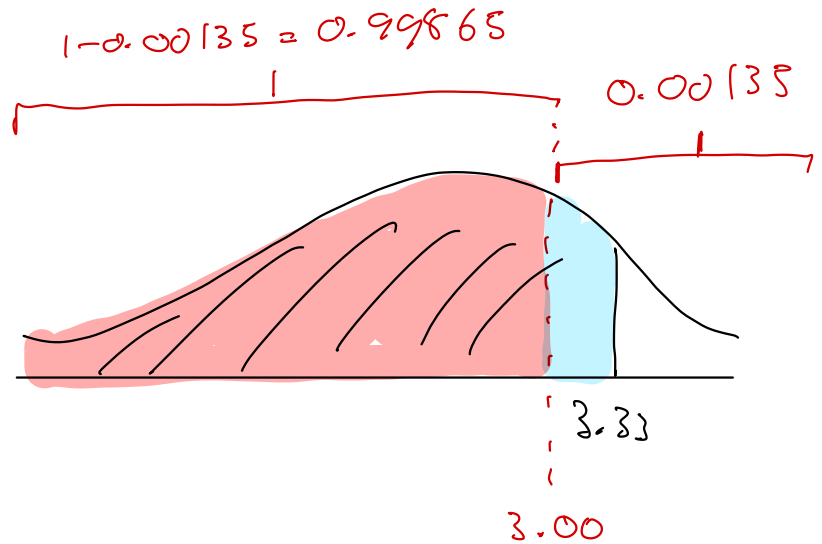
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\sim N(\mu_{\bar{x}} = 75, \sigma_{\bar{x}}^2 = 9)$$

$$P(\bar{x} \leq 85)$$

$$= P(z \leq 3.33)$$

$$= P(z \leq 3.33) > 0.99865$$



$$\begin{aligned} &\frac{R}{> pnorm(3.33)} \\ &> 0.9996 \end{aligned}$$

at least 99.865% sure an avg claim will not exceed \$85

# Solution

The Central Limit Theorem says that, in spite of the skewness of the population distribution, the average loss among 10,000 policies will be approximately  $N(75; \frac{300}{\sqrt{10000}}) = N(75; 3)$ . Now

$$P(\bar{X} > 85) \approx P\left(Z > \frac{85 - 75}{3}\right) = P(Z > 3.33) = \underline{1 - 0.9996} = 0.0004$$

We can be about 99.96% certain that average losses will not exceed \$85 per policy.

A freight elevator can transport a maximum of 9800 kg. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean 205 kg and standard deviation 15 kg. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

# Example

A manufacturer of automobile batteries claims that the distribution of the lengths of life of its best battery has a mean of 54 months and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 50 of the batteries and subjecting them to tests that estimate the battery's life.

- a) Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.
- b) Assuming that the manufacturer's claim is true, what is the probability that the consumer group's sample has a mean life of 52 or fewer months?



## Solution a)

We can use the Central Limit Theorem to deduce that the sampling distribution for a sample mean lifetime of 50 batteries is approximately Normally distributed. Furthermore, the mean of this sampling distribution ( $\mu_{\bar{X}}$ ) is 54 months according to the manufacturer's claim. Finally, the standard deviation of the sampling distribution is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485 \text{ month}$$

## Solution b)

If the manufacturer's claim is true, the probability that the consumer group observes a mean battery life of 52 or fewer months for its sample of 50 batteries is given by

$P(\bar{X} \leq 52)$ , where  $\bar{X}$  is Normally distributed,  $\mu_{\bar{X}} = 54$  and  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485$ . Hence,

$$\begin{aligned} P(\bar{X} \leq 52) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{52 - 54}{0.8485}\right) \\ &\approx P(Z \leq -2.35710076605775) \\ &\approx P(Z \leq -2.36) \end{aligned}$$

(from Table)  
= 0.0091