MATH 4720 / MSCS 5720

Instructor: Mehdi Maadooliat

Chapter 4 (Part B)

Department of Mathematics, Statistics and Computer Science

CHAPTER 4 (PART B)

- Probability
- Random process, Event, Sample space
- Set theory review
- Conditional Probability and Independence
- Bayes' Theorem
- Law of total probability
- Random Variables
  - Discrete
    - Binomial
    - Poisson
  - Continuous
    - Normal
- Normal Approximation to $Binomial(n, \pi)$
- Sampling Distribution

CONTINUOUS RANDOM VARIABLES

- A **continuous random variable** can take on values from an entire interval of the real line.
- The **probability density function** (pdf) of a continuous random variable, X, is a function $f(x)$ such that for $a < b$:

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

- The **cumulative distribution function** (cdf) of X is defined as

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

SOME RELATIONSHIPS

- **What is the relationship between f and F ?**

- $P(a \leq X \leq b) = F(b) - F(a)$

- $$\begin{aligned} P(X = a) &= P(a \leq X \leq a) \\ &= F(a) - F(a) = 0 \end{aligned}$$

REQUIREMENTS OF A PDF

- **A pdf must satisfy the following two requirements:**

$$f(x) \geq 0 \text{ for all } x$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- **A density curve shows the likelihood of a random variable at all possible values.**

- **The area under the curve and above any range of values on the horizontal axis is the proportion of all observations that fall in that range.**

SOME COMMON CONTINUOUS DISTRIBUTIONS:

- Uniform
- Exponential
- Gamma
- Weibull
- Beta
- Cauchy

- Normal (μ=mean, σ^2=variance)
- t (ν=df)
- Chi-Square (ν=df)
- F (ν_1=df_1, ν_2=df_2)

PIPELINE EXAMPLE

- **A pipeline is 100 miles long and every location along the pipeline is equally likely to break**
- **Let X be the distance measured in miles from the pipeline origin where a break occurs**
- **What is the pdf for X ?**
- **What is $P(30 \leq X \leq 50)$?**

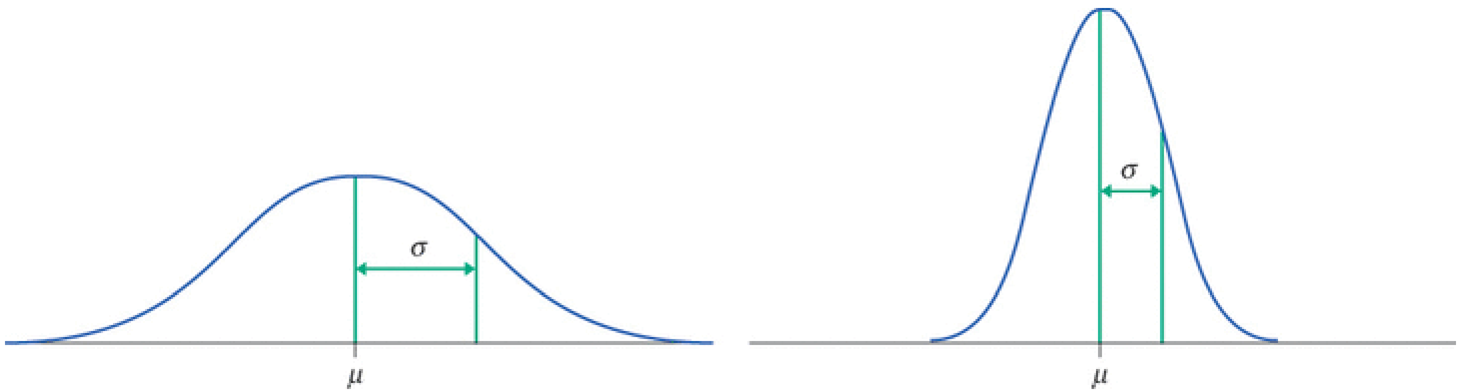
NORMAL DISTRIBUTION

- **The normal distribution, $N(\mu, \sigma^2)$, has a pdf given by**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

- **The normal distribution is always bell shaped.**
- **The normal distribution is defined in terms of its mean and variance (or standard deviation).**
- **Normal calculator**
- **Z-table (“D2L > Useful Links > Z, T and Chi^2 Tables”)**
  - $P(Z \leq z)$, **where** Z **is a standard Normal,** $Z \sim N(\mu = 0, \sigma^2 = 1)$.
- **TI-84 Calculator:**
  - **normalcdf**(a, b, μ, σ^2)**=**$P(a \leq X \leq b) = \int_a^b f(x)dx$

THE NORMAL DISTRIBUTION



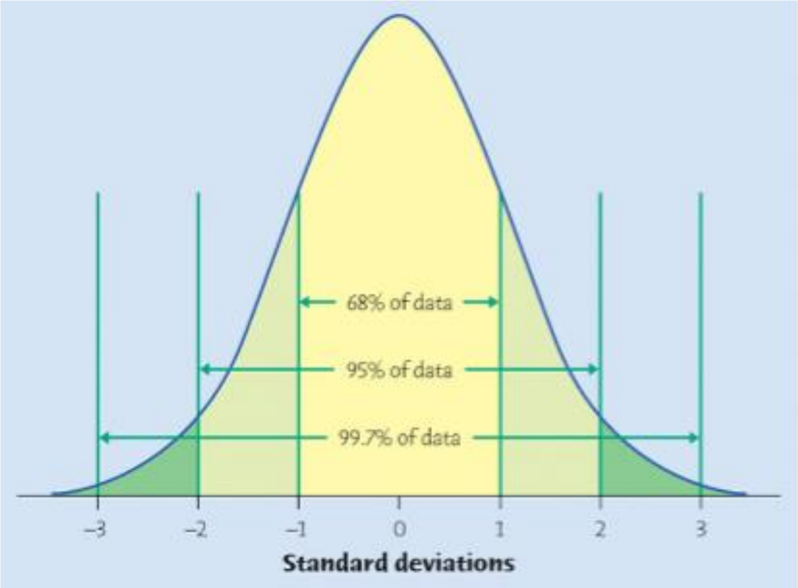
- Two Normal curves, showing the mean μ and standard deviation σ.

- The mean of a Normal distribution is at the center of the symmetric Normal curve.
- The standard deviation is the distance from the center to the change-of-curvature points on either side

- Mean and Standard deviation of Normal distribution

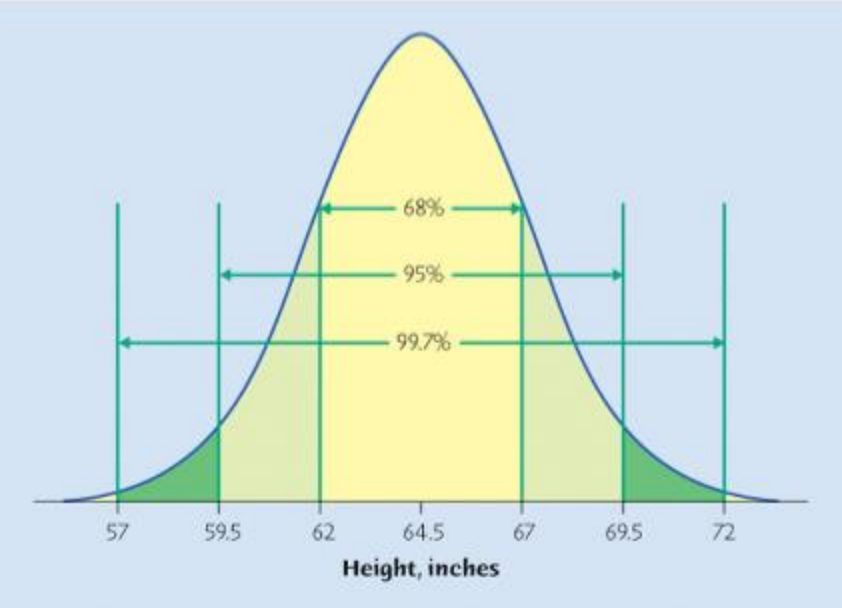
FLASHBACK TO THE 68–95–99.7 RULE (A.K.A. THE EMPIRICAL RULE)

- **In the Normal distribution with mean μ and standard deviation σ :**
  - **Approximately 68% of the observations fall within σ of the mean μ.**
  - **Approximately 95% of the observations fall within 2σ of μ.**
  - **Approximately 99.7% of the observations fall within 3σ of μ.**



EXAMPLE: HEIGHTS OF YOUNG WOMEN

- **The distribution of heights of young women aged** 18 **to** 24 **is approximately Normal with mean** $\mu = 64.5$ **inches and standard deviation** $\sigma = 2.5$ **inches. Next figure applies the** $68\text{–}95\text{–}99.7$ **rule to this distribution.**



- **Middle 95% of the heights?**

$$\mu - 2\sigma = 64.5 - (2)(2.5) = 64.5 - 5 = 59.5$$

$$\mu + 2\sigma = 64.5 + (2)(2.5) = 64.5 + 5 = 69.5$$

- **The** $68\text{–}95\text{–}99.7$ **rule applied to the distribution of heights among young women aged** 18 **to** 24**, with** $\mu = 64.5$ **inches and** $\sigma = 2.5$ **inches.**

STANDARDIZING AND Z-SCORES

- If X is an observation from a distribution that has mean μ and standard deviation σ, the **standardized value** of x is:

$$z = \frac{x - \mu}{\sigma}$$

- This standardized value is called a z-**score**.

- A z-score tells us how many standard deviations the original observation x falls away from the mean, and in which direction.
  - Observations larger than the mean have positive z-scores.
  - Observations smaller than the mean have negative z-scores.

STANDARD NORMAL DISTRIBUTION

- The standard Normal distribution is the Normal distribution $N(0,1)$ with mean 0 and standard deviation 1

- If a variable X has a Normal distribution with mean μ and standard deviation σ, then the standardized variable:

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

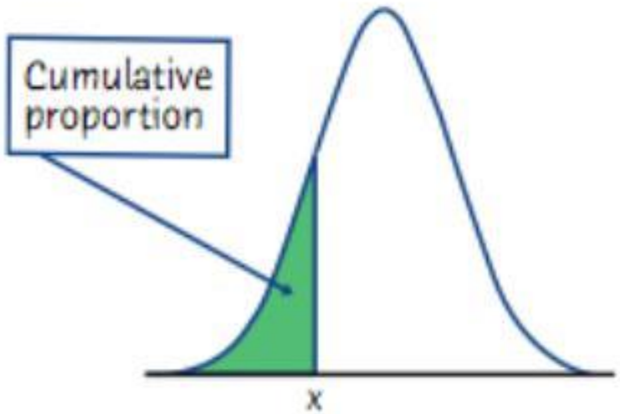
has the standard normal distribution.

EXAMPLE:
LENGTH OF HUMAN PREGNANCIES

- The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.
  - Let X be the length (in days) of a random pregnancy, the distribution of X is
  - A pregnancy was 250 days long, what is its z-score?
  - A z-score is 1.5, what is the corresponding pregnancy length?
  - What percent of babies are born after 8 months (240 days) or more of gestation from conception?

FINDING NORMAL PROBABILITIES

- The **cumulative probability** for a value x in a distribution is the proportion of observations in the distribution that lie at or below x.



- Normal calculator

- Z-table ("D2L > Useful Links > Z, T and Chi^2 Tables")
  - $P(Z \leq z)$, where Z is a **standard** Normal, $Z \sim N(\mu = 0, \sigma^2 = 1)$.

- TI-84 Calculator:
  - normal**cdf**$(-\infty, x, \mu, \sigma) = P(X \leq x)$

FINDING NORMAL PROBABILITIES WITH Z-TABLES

1. **State the problem** in terms of the observed variable x.

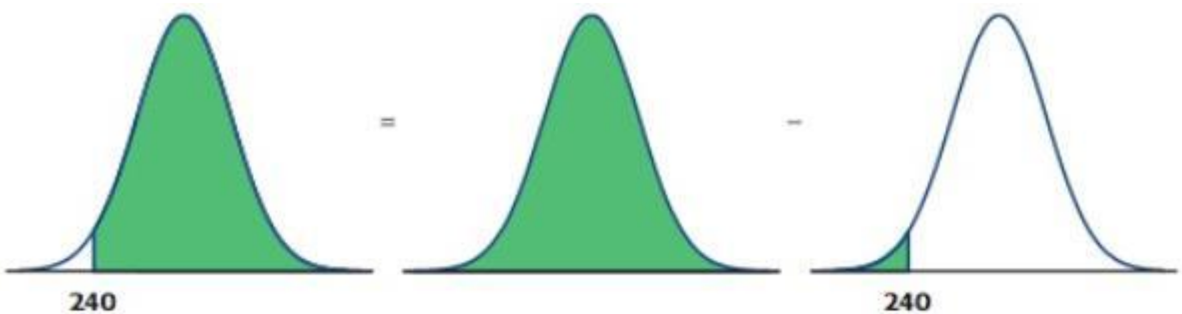
2. **Draw a picture** that shows the proportion you want in terms of cumulative proportions.

3. **Standardize x** to restate the problem in terms of a standard Normal variable z.

4. **Use Z-Tables** and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve.

LENGTH OF HUMAN PREGNANCIES (CONT'D)

- **What percent of babies are born after 8 months (240 days) or more of gestation from conception?**



area right of 240 = total area − area left of 240

= 1 − 0.0521 = 0.9479

TIPS ON FINDING NORMAL PROBABILITIES

- **Calculate** $z = \dfrac{(x - \mu)}{\sigma}$

- **Less than:**
  - $P(X < x) = P(Z < z)$

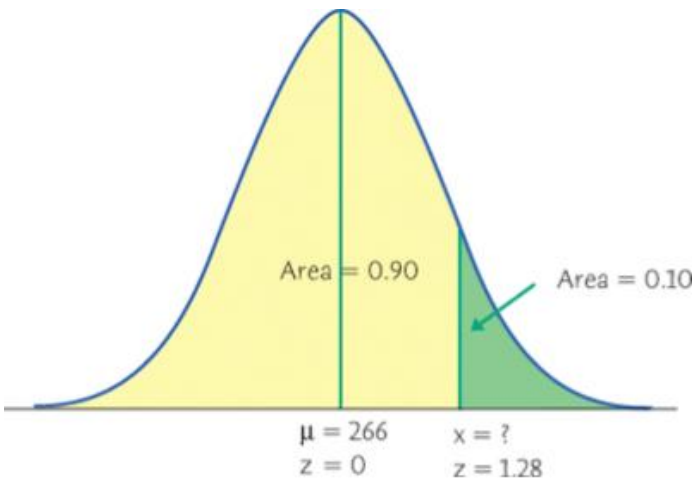
- **Greater than:**
  - $P(X > x) = P(Z > z) = 1 - P(Z < z)$

- **Between two numbers:**
  - $P(a < X < b) = P(z_a < Z < z_b) = P(Z < z_b) - P(Z < z_a)$

- **Outside of two numbers:**
  - $P(X < a \text{ OR } X > b) = P(Z < z_a \text{ OR } Z > z_b)$
    $= P(Z < z_a) + 1 - P(Z < z_b)$

LENGTH OF HUMAN PREGNANCIES (CONT'D)

- **How long are the longest 10% of pregnancies?**
  - **Step 1. State the problem and draw a picture.**
  - **Step 2. Find the cumulative probability related to the prob./proportion given.**
  - **Step 3. Find the z-score in z-tables. It is the entry closest to the cumulative prob. In this case, z = .**
  - **Step 4. Transform z back to the original x scale using the formula**



$$x = \mu + z\sigma$$

CHECKING NORMALITY: NORMAL QUANTILE PLOTS

- A Normal quantile plot (QQ plot) consists of a plot of the ordered observed data on the vertical axis and the z-scores associated with order of the observations on the horizontal axis.
- For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on. Next note that $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point.
- If the distribution of the data is close to a Normal distribution, the plotted points on a Normal quantile plot will lie close to a straight line.

NORMAL PROBABILITY PLOT

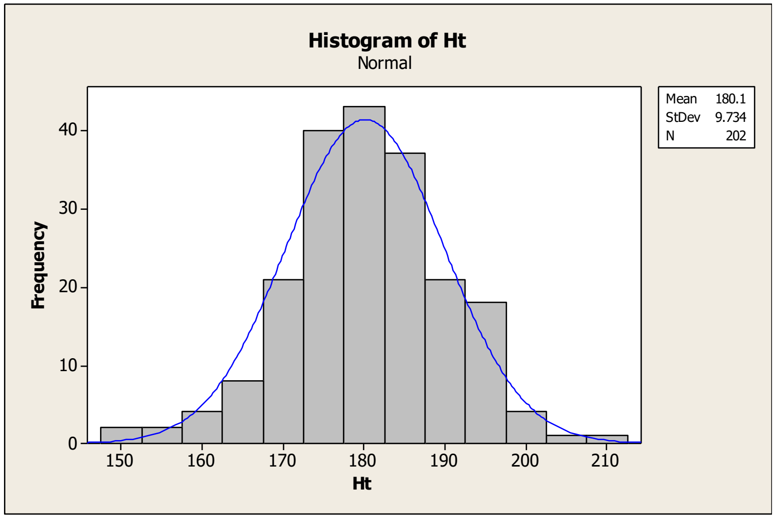
- **Rank the data from the lowest to the highest.**

$$y_{(1)} < y_{(2)} < \cdots < y_{(n)}$$

Empirical cumulative probability

i	Data Values	$\frac{i-0.5}{n}$	Normal Quantiles
1	$y_{(1)}$	**0.5/n**	q_1
2	$y_{(2)}$	**1.5/n**	q_2
.	.	.	.
.	.	.	.
.	.	.	.
n	$y_{(n)}$	**(n-0.5)/n**	q_n

AIS EXAMPLE
(HEIGHT - SUM OF SKIN FOLD)



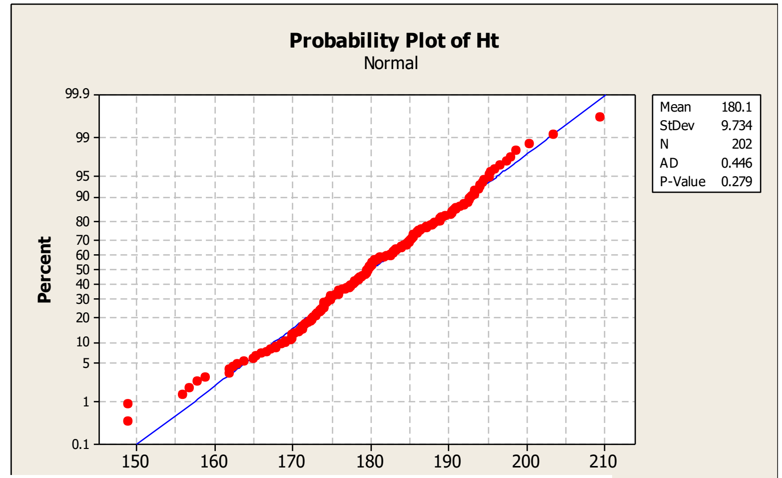
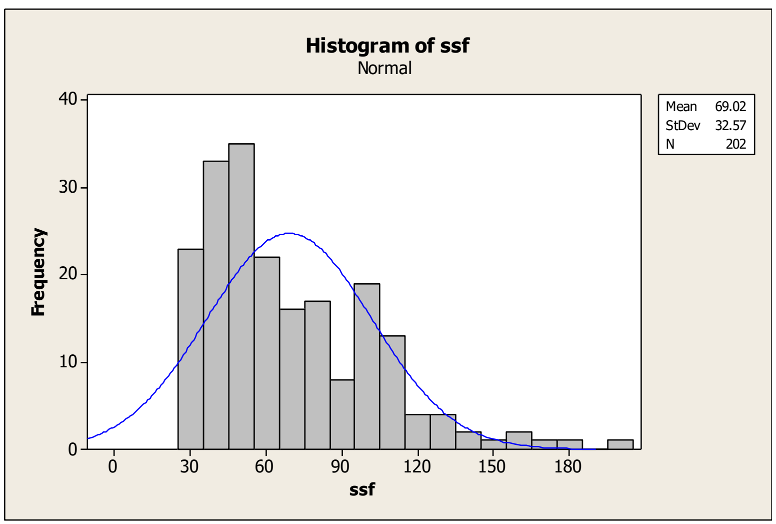
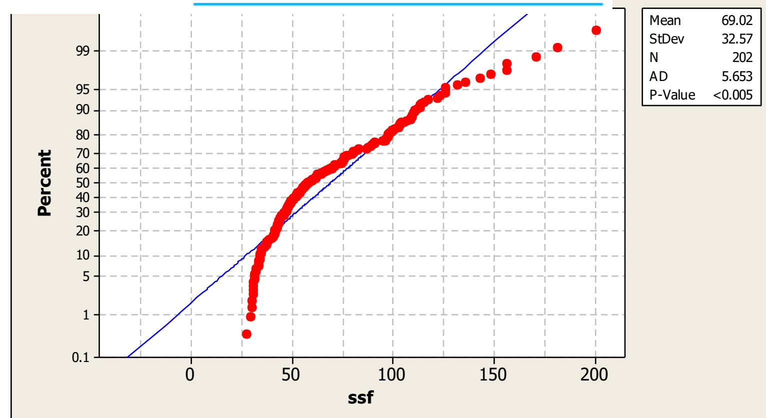


TABLE 4.13
Criteria for assessing fit of normal distribution

p-value	Assessment of Normality
$p < .01$	Very poor fit
$.01 \leq p < .05$	Poor fit
$.05 \leq p < .10$	Acceptable fit
$.10 \leq p < .50$	Good fit
$p \geq .50$	Excellent fit





BINOMIAL
MEAN AND STANDARD DEVIATION

- **If a count X has the binomial distribution with number of observations n and probability of success π, the mean and standard deviation of X are**

$$\mu = n\pi$$
$$\sigma = \sqrt{n\pi(1-\pi)}$$

NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS

- **Suppose that a count X has the binomial distribution with n observations and success probability π.**

- **When n is large, the distribution of X is approximately Normal, with mean = $n\pi$ and variance = $n\pi(1-\pi)$.**

- **As a rule of thumb, we will use the Normal approximation when n is so large that:**

$$n\pi \geq 5 \text{ and } n(1-\pi) \geq 5$$

- **Normal Approximation to Binomial Applet**

EXAMPLE: OVERWEIGHT AMERICANS

- **Nearly 60% of American adults are either overweight or obese, according to the U.S. National Center for Health Statistics. Suppose that we take a random sample of 2500 adults. What is the probability that 1520 or more of the sample are overweight or obese?**

- **Because there are almost 225 million adults, we can take the weights of 2500 randomly chosen adults to be independent. So the number in our sample who are either overweight or obese is a random variable X having the binomial distribution with $n = 2500$ and $\pi = 0.6$. To find the probability that at least 1520 of the people in the sample are overweight or obese, we must add the binomial probabilities of all outcomes from $X = 1520$ to $X = 2500$.**

OVERWEIGHT AMERICANS (CONT'D)

- **Binomial Calculator**

- **Probability distribution for the binomial model $n = 2500$ and $\pi = 0.6$, displayed graphically. The height of each bar represents the probability for X when it takes a value on the horizontal axis. Notice how the shape of this binomial probability distribution closely resembles a Normal curve.**

$$\mu = n\pi = (2500)(0.6) = 1500$$
$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(2500)(0.6)(0.4)} = 24.49$$

- **Normal Calculator**

- **The Normal approximation 0.2071 differs from the exact answer 0.2131 by only 0.006**

SAMPLING DISTRIBUTIONS

MARQUETTE UNIVERSITY
Be The Difference.

Review:

- A **parameter** is a number that describes the population.
  - Ex: population mean, population variance, population proportion

- A **statistic** is a number that can be computed from the sample data without making use of any unknown parameters. In practice, we use a statistic to estimate an unknown parameter.
  - Ex: sample mean, sample variance, sample proportion

- The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

THE SAMPLE MEAN AND THE POPULATION MEAN

MARQUETTE UNIVERSITY
Be The Difference.

- Sampling applet
- The sample mean $\bar{X}$ is a good estimate of the population mean μ.

- Means of random samples are **less variable** than individual observations.
- Means of random samples are **more Normal** than individual observations.

- **The Law of Large Numbers**: If we keep on taking larger and larger samples, the statistic $\bar{X}$ is guaranteed to get closer and closer to the population mean μ.

SAMPLING DISTRIBUTIONS OF THE SAMPLE MEAN

- **Population distribution: mean μ and standard deviation σ**
- **Sample size : n**
- **Important characteristics of the sampling distribution of the sample mean $\bar{X}$:**
- **The mean of the sampling distribution of the sample mean: $\mu_{\bar{X}}$**
- **The standard deviation of the sampling distribution of the sample mean is $\sigma_{\bar{X}}$.**
- **Fact:** $\mu_{\bar{X}} = \mu$ **and** $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$ **(also known as standard error of $\bar{X}$)**

SAMPLING DIST. OF THE SAMPLE MEAN (I)

- Sampling applet

- Population distribution: mean μ and standard deviation σ

- A random sample: sample size n, sample mean $\bar{X}$

- If the population distribution is Normal, the distribution of the mean of a random sample is also a Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

EXAMPLE: POTASSIUM IN THE BLOOD

- **There is variation both in the actual potassium level and in the blood test that measures the level. Judy's measured potassium level varies according to distribution $Normal(3.8, 0.04)$. A patient is classified as hypokalemic if the potassium level is below 3.5.**

1. **If a single potassium measurement is made, what is the probability that Judy is diagnosed as hypokalemic?**

POTASSIUM IN THE BLOOD (CONT'D)

- Distribution is $Normal(3.8, 0.04)$

- If measurements are made instead on 4 separate days and the mean result is compared with the criterion 3.5, what is the probability that Judy is diagnosed as hypokalemic?

SAMPLING DIST. OF THE SAMPLE MEAN (II)

MARQUETTE UNIVERSITY
Be The Difference.

- Sampling applet
- Population distribution: mean μ and standard deviation σ
- A random sample: sample size n, sample mean $\bar{X}$

- **Central Limit Theorem** : If the population distribution is **NOT** normal, as the sample size, n, becomes **large** the distribution of the mean of a random sample **converges** to a Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

- A sample size of at least 30 is typically required to use the **CLT**.
- The amazing part of this theorem is that it is true **regardless** of the form of the underlying distribution.

EXAMPLE 2: VACCINE FOR HIV

- On the average, HIV patients survive for 5 years after being diagnosed. A new vaccine is developed to fight the virus. In a clinical trial, 50 HIV patients were given this vaccine, and the average survival years for this sample was more than 5.6 years. Compute the probability that the sample average is more than 5.6 years assuming the population mean of 5 years and the population standard deviation of 0.6.
    - $\bar{Y} \approx N(5, 0.085^2)$
    - $P(\bar{Y} > 5.6) = normcdf(5.6, \infty, 5.0,0.085)$
    - $P(\bar{Y} > 5.6) = 8.3955 * 10^{-13}$
- What does this imply?