Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 20, 2019

Topics

Objectives

Objectives:

- Review key concepts from MTH 3210.

Review: Random Variables and Expected Values

- A ***random variable*** (***rv***) is a variable whose value is determined by chance.

Review: Random Variables and Expected Values

- A ***random variable*** (***rv***) is a variable whose value is determined by chance.
- The ***probability distribution*** of a rv indicates:
  1. The **values** that the variable might take.
  2. The **probabilities** of those values.

Review: Random Variables and Expected Values

- A ***random variable*** (***rv***) is a variable whose value is determined by chance.
- The ***probability distribution*** of a rv indicates:
  1. The **values** that the variable might take.
  2. The **probabilities** of those values.
- **Probability distributions** are represented by:
  - ***Probability mass functions*** (or ***pmfs***) (**discrete** rvs).
  - ***Probability density functions*** (or ***pdfs***) (**continuous** rvs).

- The ***mean*** (or ***expected value***) of a rv X, denoted μ (or $E(X)$), is

- The ***mean*** (or ***expected value***) of a rv X, denoted μ (or $E(X)$), is

Mean (or Expected Value):

$$\mu = E(X) = \begin{cases} \sum_i x_i p(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

- The ***mean*** (or ***expected value***) of a rv X, denoted μ (or $E(X)$), is

Mean (or Expected Value):

$$\mu = E(X) = \begin{cases} \sum_i x_i p(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

- μ measures the **center of the distribution** and represents the **long-run average** of X.

- The ***variance*** of X, denoted $\boldsymbol{\sigma^2}$ (or $\boldsymbol{V(X)}$) is

Variance:

$$\begin{aligned}\sigma^2 &= E\left((X-\mu)^2\right) \\ &= \begin{cases} \sum_i (x_i-\mu)^2 p(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}\end{aligned}$$

- The ***standard deviation*** of X, denoted $\boldsymbol{\sigma}$ (or $\boldsymbol{SD(X)}$), is the square root of the variance:

- The ***standard deviation*** of X, denoted $\boldsymbol{\sigma}$ (or $\boldsymbol{SD(X)}$), is the square root of the variance:

Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

- The ***standard deviation*** of X, denoted $\boldsymbol{\sigma}$ (or $\boldsymbol{SD(X)}$), is the square root of the variance:

Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

- $\boldsymbol{\sigma^2}$ and $\boldsymbol{\sigma}$ both measure the **spread of the probability distribution** away from μ.

- The ***standard deviation*** of X, denoted $\boldsymbol{\sigma}$ (or $\boldsymbol{SD(X)}$), is the square root of the variance:

Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

- $\boldsymbol{\sigma^2}$ and $\boldsymbol{\sigma}$ both measure the **spread of the probability distribution** away from μ.

  $\boldsymbol{\sigma}$ is measured in the **same units** as $\boldsymbol{X}$, and represents the size of a **typical deviation** of X away from μ.

Proposition

Mean and Variance of a Constant: If a is any constant, then

- $E(a) = a$

Proposition

Mean and Variance of a Constant: If a is any constant, then

- $E(a) = a$
- $V(a) = 0$ and $SD(a) = 0$

Proposition

Mean and Variance of a Linear Function of X: If X is *any* random variable whose mean and variance are μ and σ^2, and a and b are any constants, then

- $E(aX + b) \;=\; a\mu + b$

Proposition

Mean and Variance of a Linear Function of X: If X is *any* random variable whose mean and variance are μ and σ^2, and a and b are any constants, then

- $E(aX + b) \ = \ a\mu + b$
- $V(aX + b) \ = \ a^2\sigma^2$ and $SD(aX + b) \ = \ |a|\sigma$

Review: The Normal Distribution

- A random variable X is said to have a ***normal*** distribution with **parameters** μ and σ, denoted $N(\mu, \sigma)$, if its pdf is

Review: The Normal Distribution

- A random variable X is said to have a ***normal*** distribution with **parameters** μ and σ, denoted $N(\mu, \sigma)$, if its pdf is

Normal pdf:

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

Review: The Normal Distribution

- A random variable X is said to have a ***normal*** distribution with **parameters** $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, denoted $N(\boldsymbol{\mu}, \boldsymbol{\sigma})$, if its pdf is

Normal pdf:

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

- $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the **mean** and **standard deviation** of the $N(\boldsymbol{\mu}, \boldsymbol{\sigma})$ distribution.

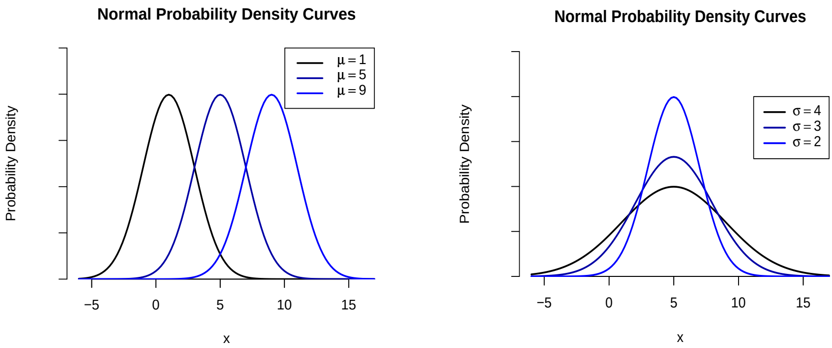


Figure: Normal distributions with different values of μ, but the same σ (left), and with the same μ, but different values of σ (right).

- The notation $\boldsymbol{X \sim N(\mu, \sigma)}$ is short for "X follows a $N(\mu, \sigma)$ distribution."

- The notation $\boldsymbol{X \sim N(\mu, \sigma)}$ is short for "X follows a $N(\mu, \sigma)$ distribution."

Proposition

Linear Function of a Normal Random Variable: If $X \sim N(\mu, \sigma)$ and we let

$$Y = aX + b,$$

where a and b are constants, then

$$Y \sim N\left(a\mu + b,\ |a|\sigma\right).$$

- The $\boldsymbol{N(0,1)}$ distribution ($\mu = 0$ and $\sigma = 1$) is called the ***standard normal*** distribution.

- The $\boldsymbol{N(0,1)}$ distribution ($\mu = 0$ and $\sigma = 1$) is called the ***standard normal*** distribution.

Proposition

Standardizing a Normal Random Variable: If $X \sim N(\mu, \sigma)$ and we let

$$Z = \frac{X - \mu}{\sigma}, \tag{1}$$

then

$$Z \sim N(0, 1).$$

- The transformation (1) from X to Z is called ***standardizing*** X, and Z is measured in ***standard units***, which are standard deviations away from the mean.

Review: Linear Combinations of Random Variables

Mean and Variance of a Linear Combination of Random Variables

- Random variables $X_1, X_2, \ldots, X_n$ are said to be ***independent*** if their values aren't influenced by each other.

Review: Linear Combinations of Random Variables

Mean and Variance of a Linear Combination of Random Variables

- Random variables $X_1, X_2, \ldots, X_n$ are said to be ***independent*** if their values aren't influenced by each other.
- $X_1, X_2, \ldots, X_n$ are said to be ***iid*** (for ***independent and identically distributed***) if they're drawn independently from a single probability distribution.

Review: Linear Combinations of Random Variables

Mean and Variance of a Linear Combination of Random Variables

- Random variables $X_1, X_2, \ldots, X_n$ are said to be ***independent*** if their values aren't influenced by each other.
- $X_1, X_2, \ldots, X_n$ are said to be ***iid*** (for ***independent and identically distributed***) if they're drawn independently from a single probability distribution.
- The term ***random sample*** will be taken to mean **iid** observations.

- For random variables $X_1, X_2, \ldots, X_n$ and any constants $a_1, a_2, \ldots, a_n$, the new random variable

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

is called a ***linear combination*** of the X_i's.

Proposition

If $X_1, X_2, \ldots, X_n$ are *any* random variables (not necessarily independent) whose means are $\mu_1, \mu_2, \ldots, \mu_n$, respectively, then for any constants $a_1, a_2, \ldots, a_n$,

$$E\left(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n\right) \;=\; a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n.$$

As a special case, if $X_1, X_2, \ldots, X_n$ are a random sample from a distribution whose mean is μ, then

$$E\left(X_1 + X_2 + \cdots + X_n\right) \;=\; \mu + \mu + \cdots + \mu \;=\; n\mu.$$

Proposition

If $X_1, X_2, \ldots, X_n$ are any *independent* random variables whose variances are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively, then for any constants $a_1, a_2, \ldots, a_n$,

$$V(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2$$

and

$$SD(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) = \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2}.$$

As a special case, if $X_1, X_2, \ldots, X_n$ are a random sample from a distribution whose variance is σ^2, then

$$V(X_1 + X_2 + \cdots + X_n) \;=\; \sigma^2 + \sigma^2 + \cdots + \sigma^2 \;=\; n\sigma^2$$

and

$$SD(X_1 + X_2 + \cdots + X_n) \;=\; \sqrt{\sigma^2 + \sigma^2 + \cdots + \sigma^2} \;=\; \sqrt{n}\sigma.$$

Linear Combinations of *Normal* Random Variables

Proposition

Suppose $X_1, X_2, \ldots, X_n$ are *independent*, with $X_i \sim N(\mu_i, \sigma_i)$. Let

$$Y \;=\; a_1X_1 + a_2X_2 + \cdots + a_nX_n \;=\; \sum_{i=1}^{n} a_iX_i$$

(where the a_i's are any constants). Then

$$Y \;\sim\; N\left(\sum_{i=1}^{n} a_i\mu_i,\;\; \sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}\right).$$

As a special case, if $X_1, X_2, \ldots, X_n$ are a random sample from a $N(\mu, \sigma)$ distribution, then

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, \ \sqrt{n}\sigma\right).$$

Review: Statistics and Sampling Distributions

Statistics

- Any numerical value computed from a **random sample** $X_1, X_2, \ldots, X_n$ is called a ***statistic***.

Review: Statistics and Sampling Distributions

Statistics

- Any numerical value computed from a **random sample** $X_1, X_2, \ldots, X_n$ is called a ***statistic***.
- The **sample mean** and **sample standard deviatin** are two important statistics:

Review: Statistics and Sampling Distributions

Statistics

- Any numerical value computed from a **random sample** $X_1, X_2, \ldots, X_n$ is called a ***statistic***.
- The **sample mean** and **sample standard deviatin** are two important statistics:

Sample Mean: The ***sample mean***, denoted $\bar{X}$, is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,. \qquad (2)$$

Sample Variance and Standard Deviation: The ***sample variance***, denoted $S^{\mathbf{2}}$, is

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

and the ***sample standard deviation***, denoted S, is

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}.$$

The Sampling Distribution of $\bar{X}$ and the Central Limit Theorem

- **Statistics** are are **random variables.**

The Sampling Distribution of $\bar{X}$ and the Central Limit Theorem

- **Statistics** are are **random variables.**
- The probability distribution of a statistic is called its ***sampling distribution***.

Proposition

If $X_1, X_2, \ldots, X_n$ are a random sample from *any* distribution (not necessarily normal) whose mean and standard deviation are μ and σ, then

$$E(\bar{X}) = \mu$$

and

$$\text{V}(\bar{X}) = \frac{\sigma^2}{n} \qquad \text{and} \qquad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Proposition

If $X_1, X_2, \ldots, X_n$ are a random sample from *any* distribution (not necessarily normal) whose mean and standard deviation are μ and σ, then

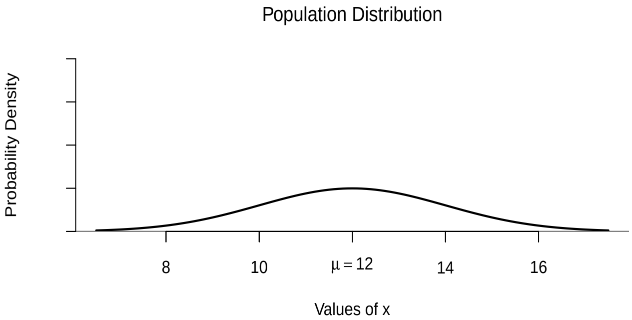
$$E(\bar{X}) = \mu$$

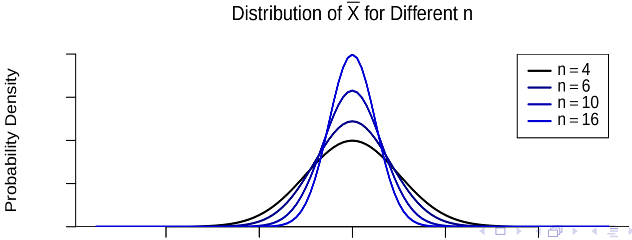
and

$$\text{V}(\bar{X}) = \frac{\sigma^2}{n} \qquad \text{and} \qquad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

This follows from the fact that $\bar{X}$ is a **linear combination** of $X_1, X_2, \ldots, X_n$.

- The standard deviation $\sigma/\sqrt{n}$ of $\bar{X}$ is sometimes call the ***standard error*** of $\bar{X}$, and represents a **typical deviation** of $\bar{X}$ away from μ.


Population Distribution
Probability Density
8
10
μ = 12
14
16
Values of x


Distribution of $\overline{X}$ for Different n
Probability Density
n = 4
n = 6
n = 10
n = 16

Proposition

Sampling Distribution of $\bar{X}$ Under Normality of the X_i's:
Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a $N(\mu, \sigma)$ distribution. Then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \tag{3}$$

and so

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

Proposition

Sampling Distribution of $\bar{X}$ Under Normality of the X_i's:
Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a $N(\mu, \sigma)$ distribution. Then

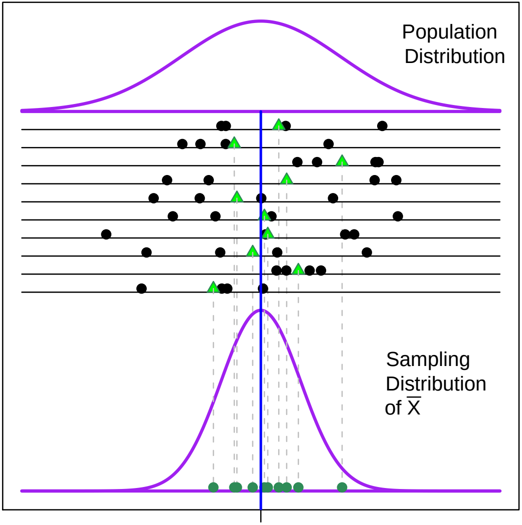
$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \tag{3}$$

and so

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

This follows from the fact that $\bar{X}$ is a **linear combination** of $X_1, X_2, \ldots, X_n$.

Population Distribution and Sampling Distribution of $\overline{X}$



Proposition

Central Limit Theorem: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from *any* distribution whose mean and standard deviation are μ and σ, with $\sigma < \infty$. Then **if n is large**,

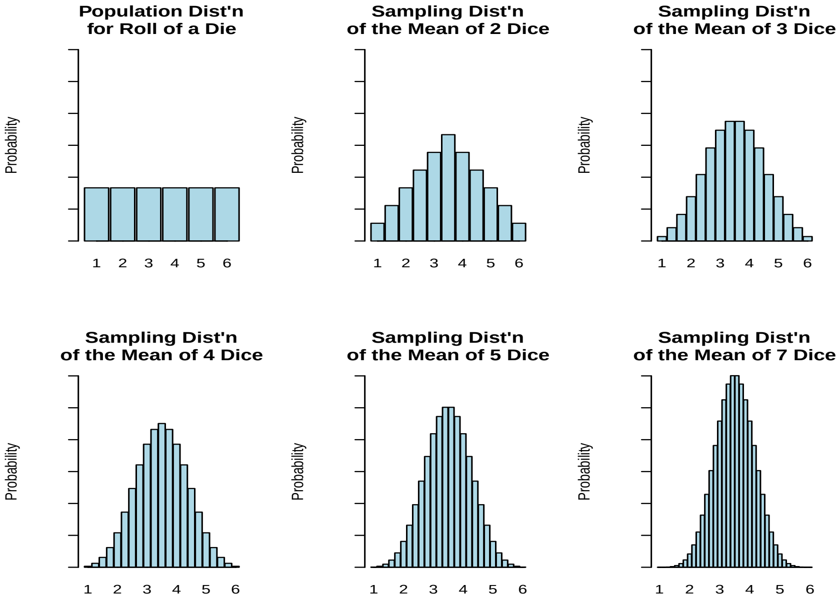
$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

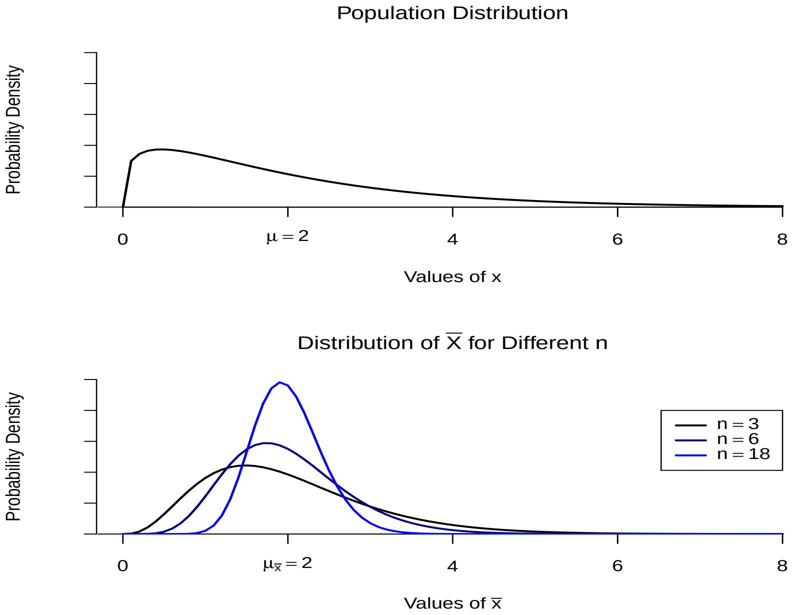
approximately, and in this case

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

approximately.

- The larger n is, the more closely the $\bar{X}$ distribution resembles the normal distribution.


Population Dist'n
for Roll of a Die
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 2 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 3 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 4 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 5 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 7 Dice
Probability
1 2 3 4 5 6


Population Distribution
Probability Density
0
$\mu = 2$
4
6
8
Values of x
Distribution of $\overline{X}$ for Different n
Probability Density
n = 3
n = 6
n = 18
0
$\mu_{\overline{x}} = 2$
4
6
8
Values of $\overline{x}$

The Law of Large Numbers

Proposition

Law of Large Numbers: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from *any* distribution whose mean and standard deviation are μ and σ, with $\sigma < \infty$. Then

$$\bar{X} \to \mu$$

as $n \to \infty$.

The Law of Large Numbers

Proposition

Law of Large Numbers: Suppose $X_1, X_2, \dots, X_n$ are a random sample from *any* distribution whose mean and standard deviation are μ and σ, with $\sigma < \infty$. Then

$$\bar{X} \to \mu$$

as $n \to \infty$.

(Each time n is increased by 1, we recompute $\bar{X}$, giving a *sequence* of $\bar{X}$ values, which get closer and closer to μ.)

Review: t Distributions

- If $X_1, X_2, \ldots X_n$ are a random sample from a $N(\mu, \sigma)$ distribution, the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \qquad (4)$$

follows a ***t distribution*** with $n - 1$ ***degrees of freedom*** (***df***), denoted $t(n-1)$.

- **Properties of t distributions**:

- **Properties of t distributions**:

  1. They're centered on 0, and resemble the $N(0, 1)$ distribution, but have heavier tails.

- **Properties of t distributions**:

  1. They're centered on 0, and resemble the $N(0,1)$ distribution, but have heavier tails.
  2. As the df increases, the t distributions approach the $N(0,1)$ distribution.

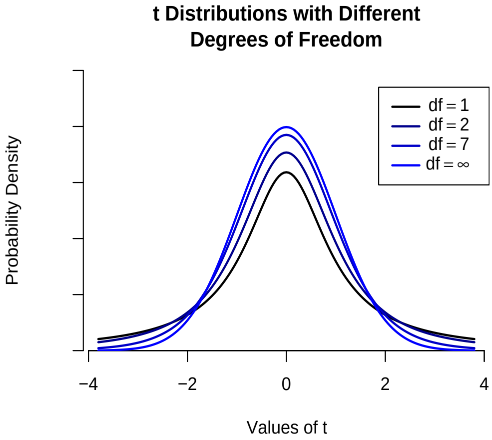


Figure: t distributions with different degrees of freedom. The t distribution with ∞ degrees of freedom is the $N(0, 1)$ curve.

- Even if a sample is from a **non-normal** distribution, the random variable (4) follows a t distribution if n is large.

- Even if a sample is from a **non-normal** distribution, the random variable (4) follows a t distribution if n is large.

Proposition

Suppose $X_1, X_2, \ldots X_n$ are a random sample from *any* distribution whose mean and standard deviation are μ and σ. Then **if n is large**,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

approximately.

- The above fact is a consequence of the facts that

- The above fact is a consequence of the facts that

  1. $S \to \sigma$ as $n \to \infty$
  2. $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$ when n is large (by the CLT)
  3. The $t(n-1)$ and $N(0,1)$ distributions are nearly identical when n is large.

Review: Confidence Interval for μ

One-Sample t CI: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose mean is μ. Then a $100(1-\alpha)\%$ ***one-sample t confidence interval*** (***CI***) for μ is

$$\bar{X} \pm t_{\alpha/2,\, n\,-\,1} \frac{S}{\sqrt{n}}, \tag{5}$$

where $t_{\alpha/2,\, n\,-\,1}$ is the $100(1-\alpha/2)$th percentile of the $t(n-1)$ distribution.

- The CI is valid if either the sample is from a **normal** population or n **is large.**

- The CI is valid if either the sample is from a **normal** population or n **is large.**
- In either case, we can be $100(1-\alpha)\%$ confident that μ will be contained in the CI.