ISIMS - UNIVERSITY OF SFAX
LSI ADBD - BACHELOR 2
2023/2024

***

DESCRIPTIVE AND INFERENTIAL STATISTICS

***

LECTURE V

THE NORMAL DISTRIBUTION AND OTHER CONTINUOUS DISTRIBUTIONS

Lecture Content

THE NORMAL DISTRIBUTION

The Normal Distribution

The normal distribution or Gauss distribution $\mathcal{N}(\mu, \sigma)$ is one of the symmetrical distributions and is described by the normal equation. The normal distribution conerns random continuous variable defined on the sample space $\mathbb{R}$.

The normal distribution's density function is defined by the following equation:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp \frac{-(x-\mu)^2}{2\sigma^2}$$

where X is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and exp is the exponential function.
The random variable X in the normal equation is called the normal random variable.

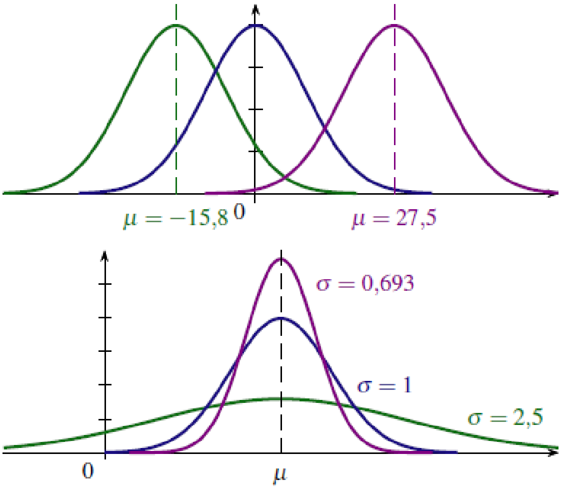
Shape of the Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation.

- The mean of the distribution determines the location of the center of the graph.
- The standard deviation determines the height and width of the graph.
- When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow.
- All normal distributions look like a symmetric, bell-shaped curve, as shown below.

The Normal Curve

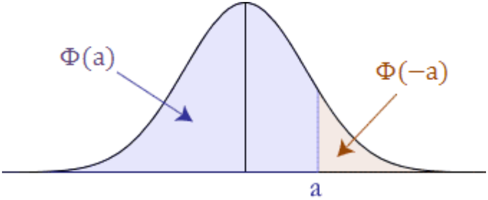
Graph of the Probability Density Function



Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is infinitesimal, considered as 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and $+\infty$ (the yellow area).
- The probability that X is less than a equals the area under the normal curve bounded by a and $-\infty$ (the blue area).



Empirical Rule and Normal Distribution

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the empirical rule or the 68-95-99.7 rule. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

Standard Normal Distribution $\mathcal{N}(0, 1)$

- The standard normal distribution is a normal distribution which has a mean of zero and a standard deviation of one, $\mathcal{N}(0, 1)$.
- The normal random variable of a standard normal distribution is called a standard score or a z-score.
- Every normal random variable X following a $\mathcal{N}(\mu, \sigma)$ can be transformed into a z-score via the following equation:

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

- **The standard normal distribution table**
  - A standard normal distribution table shows the cumulative distribution associated with a particular z-score.
  - Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place.
  - The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table.

Standard Normal Distribution Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
...	...	...	...	...	...	...	...	...	...	...
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
...	...	...	...	...	...	...	...	...	...	...
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Example 1: A section of the standard normal table is reproduced below. To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, $P(Z < -1.31) = 0.0951$.

Standard Normal Distribution Table

Your table gives you the cumulative $P(Z < a) = \Phi(a)$. Then to

- Find $P(Z > a)$: The probability that a standard normal distribution will be greater than a. Here it is

$$P(Z > a) = 1 - P(Z < a) = 1 - \Phi(a)$$

- Find $P(a < Z < b)$. The probability that a standard normal random variable lies between two values is also easy to find.

$$P(a < Z < b) = P(Z < b) - P(Z < a)$$

Example 2: Suppose we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table (see above), we find that
$P(Z < -1.20) = 0.1151$; and $P(Z < -1.40) = 0.0808$. Therefore,
$P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) =$
$0.1151 - 0.0808 = 0.0343$

The Normal Distribution as a Model for Measurements

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

1. Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier:

$$z = (X - \mu)/\sigma$$

2. Find probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators.

Normal probability Calculations - Exercises

Exercise 1: Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed.)

Exercise 2: An average light bulb manufactured by the ACME Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an ACME light bulb will last at least 365 days?

Exercise 3: Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

To find percentiles (Inverse the Normal distribution)

Given the cumulative probability we find its z-sore and then the x-score by using the equation

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

Example 3: To find the smallest IQ of the 5% smartest people, Plot the z-curve. The area below this IQ score is 95%. This gives from the table a z-score= 1.645 (i.e. this person has to be at least 1.645 standard deviation above average).
Hence, the min IQ score for the top 5% of the data equals

$$X = 100 + 1.645(10) = 164.5$$

$$\sim \circ \sim$$

With Excel:
To find cumulative probability:
NORMDIST(x, mean, s, cumulative (type true))
3mm To find percentile:
NORMINV(p, mean, s)

OTHER CONTINUOUS DISTRIBUTIONS

The Uniform Distribution

A continuous random variable X has a uniform distribution if its density is constant on a finite interval $[a, b]$, i.e:

$$f(x) = \begin{cases} k & if \quad x \in [a, b] \\ 0 & else \end{cases}$$

We write $X \sim \mathcal{U}([a, b])$. The density is:

$$f(x) = \frac{1}{b - a}, \quad \forall x \in [a, b]$$

Exercise 1: Show this result and plot the density graph.
Exercise 2: Find the cumulative distribution function and plot it in a graph.

The Uniform Distribution

Probability of an Interval: The probability of an interval $[x_1, x_2]$ in $[a, b]$ is proportional to its magnitude.

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = \frac{1}{b-a}\int_{x_1}^{x_2} dx = \frac{x_2 - x_1}{b-a}$$

The Expected Value:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{b-a}\int_a^b xdx = \frac{b+a}{2}$$

The Variance:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^2 + ab + a^2}{3}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{(b-a)^2}{12}$$

The Exponential Distribution

The Exponential Distribution with parameter $\theta > 0$ concerns a positive random variable with density:

$$f(x) = \begin{cases} \theta e^{-\theta x} & \textit{if} \quad 0 \leq x \\ 0 & \textit{if} \quad x < 0 \end{cases}$$

We write $X \sim \mathcal{E}(\theta)$, with X taking its values in $\mathbb{R}_+$

The exponential distribution is often used to model the probability distribution of random variables that express duration.

Exercise 1: Find the Cumulative Distribution Function.
Exercise 2: Using integration by parts technique, show that $E(X) = 1/\theta$ and that $V(X) = 1/\theta^2$.

EMPIRICAL DISTRIBUTIONS

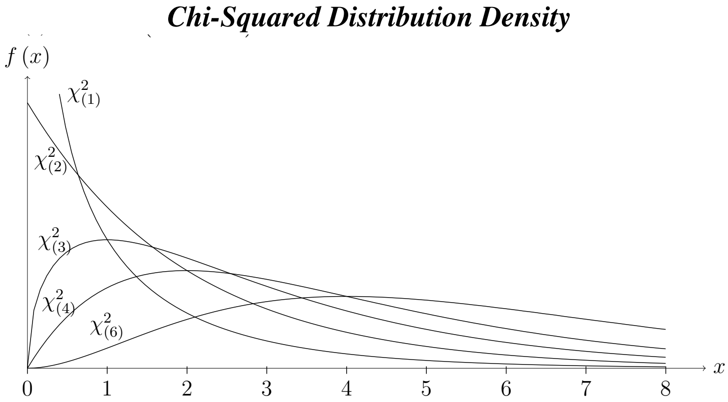
The Chi-Squared Distribution

If $X_1, ..., X_n$ is a random variable sample such that
$X_i \sim^{i.i.d} \mathcal{N}(0,1) \quad \forall i = 1, ..., n$ then $\sum_{i=1}^{n} X_i^2 = X \sim \mathcal{X}^2_{(n)}$
We denote by $k_{(p,n)}$ the p-quantile of $\mathcal{X}^2_{(n)}$; i.e. the value such that: If $X \sim \mathcal{X}^2_{(n)}$, Then $Pr(X \leq k_{(p,n)}) = p$.

Properties:

- If $X \sim \mathcal{X}^2_{(n)}$, Then $E(X) = n \quad$ et $V(X) = 2n$.
- If $X_1 \sim \mathcal{X}^2_{(n_1)}$; $X_2 \sim \mathcal{X}^2_{(n_2)}$ et $X_1 \perp X_2$, Then $(X_1 + X_2) \sim \mathcal{X}^2_{(n_1+n_2)}$.
- If $X = X_1 + X_2$, $X \sim \mathcal{X}^2_{(n)}$ and $X_1 \sim \mathcal{X}^2_{(p)}$ with $p < n$, Then $X_2 \sim \mathcal{X}^2_{(n-p)}$ and $X_1 \perp X_2$.

The Chi-Squared Distribution - Graph and table



The Student Distribution

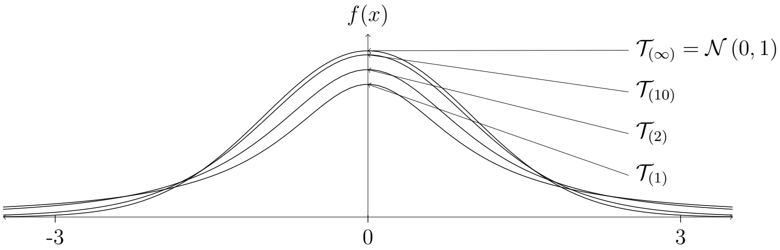
If $Z \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{X}^2_{(n)}$ and $Z \perp X$, Then $T = \frac{Z}{\sqrt{X/n}} \sim \mathcal{T}_{(n)}$.

We denote $t_{(p,n)}$ the p-quantile of $\mathcal{T}_{(n)}$; i.e. the value such that: If $T \sim \mathcal{T}_{(n)}$, Then $Pr(T \leq t_{(p,n)}) = p$.

Properties:

- If $T \sim \mathcal{T}_{(n)}$, then $E(T) = n \quad$ and $V(T) = \frac{n}{n-1}$.
- The Student distribution converges to $\mathcal{N}(0,1)$ when the degrees of freedom becomes large.

The Student Distribution - Graph and table



The Fisher Distribution

If $X_1 \sim \mathcal{X}^2_{(n_1)}$ and $X_2 \sim \mathcal{X}^2_{(n_2)}$ and $X_1 \perp X_2$, Then $F = \frac{X_1/n_1}{X_2/n_2} \sim \mathcal{F}_{(n_1,n_2)}$.
We denote $f_{(p,n_1,n_2)}$ the p-quantile of $\mathcal{F}_{(n_1,n_2)}$; i.e. the value such that:
If $F \sim \mathcal{F}_{(n_1,n_2)}$, Then $Pr(F \leq f_{(p,n_1,n_2)}) = p$.

The Fisher Distribution - Graph and Table