LECTURE 6: PERCENTILES, NORMAL DISTRIBUTIONS

AP STATS

ABSTRACT. We first review cumulative distributions to help determine percentiles in an ordered list. We then relate z-scores to percentiles in the case of normal distributions.

1. Cumulative Frequencies and Cumulative Distributions

Given a random variable X and a PMF/PDF f defined on X, we can consider *another distribution* called the **cumulative distribution function**, which sums all of the probabilities up to a given point.

Definition. The **cumulative distribution function** (CDF) is the function defined as

$$CDF(x) = \int_{t=-\infty}^{x} f(t), \quad f = PDF/PMF$$

The cumulative distribution function has the following properties:

- Domain = all real numbers $\mathbb{R}$
- Range of outputs = $[0, 1]$
- CDF is *monotonically increasing* (if $x \leq y \implies CDF(x) \leq CDF(y)$)

Another way of defining the CDF is by

$$CDF(x) = Prob(X \leq x).$$

That is, the CDF will tell you the probability *up to a certain point* x (you get to pick this point!).

Example 1. From **Question 1**, we have $X = \{1, 2, 5\}$ and $p(x) = \dfrac{1}{8}$ as our probability function.

Let's find the values of the CDF:

x	1	2	5
$p(x)$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{5}{8}$
$CDF(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{8}{8}$

Note that we set the $CDF(x) = 0$ for any $x < 1$ in this example so that the CDF starts life at 0 and then concludes its final output to be $8/8 = 1$.

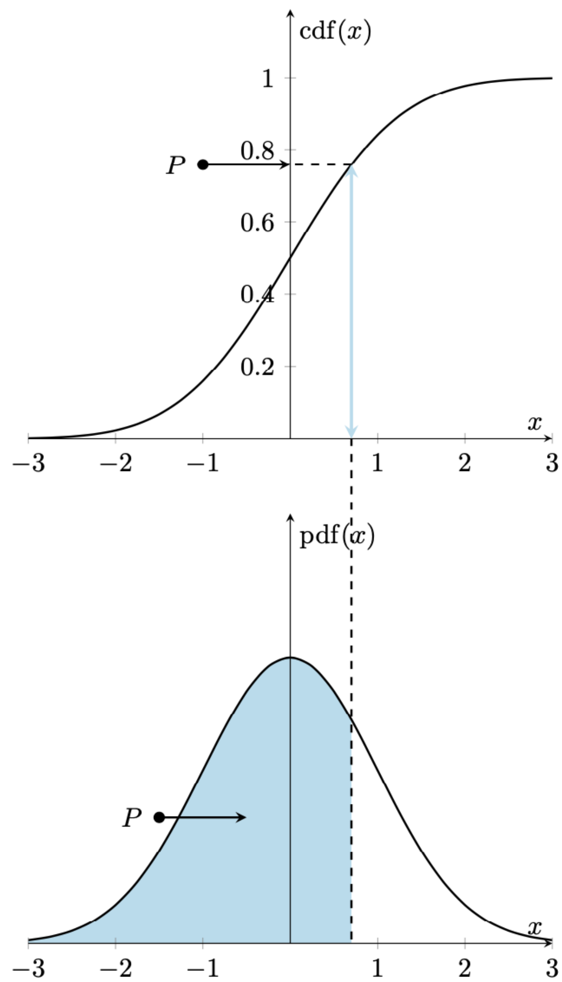


FIGURE 1. Relationship between CDF and PDF

Example 2. If we set $X = [2, 2023]$ and let $f(x) = \dfrac{1}{2021}$ to be our (uniform) probability function, then we can evaluate the CDF at any given point between 2 and 2023. So if we were interested in $x = 2020$, then

$$CDF(x = 2020) = [\text{area from } x = 2 \text{ to } x = 2020] = \frac{2018}{2021}$$

The graphic displaying different CDF's showcases their shared properties and how tweaking the mean or standard deviation changes how quickly we increase[1]:

[1] Image is from here

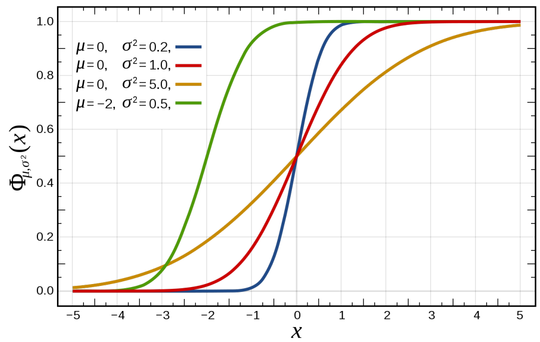


FIGURE 2. Different CDFs

1.1. Cumulative Frequencies.

Since the CDF is defined in terms of the PDF, we get that $CDF(x) \leq 1$ for all inputs x. However, if we are given a *frequency table*, we can convert these frequencies into **relative frequencies** to get the probabilities, but we can also consider the **cumulative frequency.**

Let's work with a particular example:

Example 3. Suppose we have weights of a number of people in kilograms (kg) with the following frequencies:

weight	**frequency**	***cumulative frequency***
$55 \leq w < 60$	2	2
$60 \leq w < 65$	3	5
$65 \leq w < 70$	12	17
$70 \leq w < 75$	20	37
$75 \leq w < 80$	13	50
$80 \leq w < 85$	10	60
$85 \leq w < 90$	5	65

2. Z-Scores

Definition. Given a quantitative variable, the **z-score** of a point x is the number

$$z_x = \frac{x-\mu}{\sigma}, \quad \mu = \text{mean}, \quad \sigma = \text{standard dev}$$

The z-score tells you how many standard deviations you are away from the mean μ.

Example 1. Suppose $\mu = 70$ years and $\sigma = 5$ years. Say that your tortoise lifespan was in the 2.3 standard deviations above the mean. How long was your tortoise lifespan?

We write $z_x = 2.3$ and then do the computations:

$$2.3 = z_x = \frac{x - 70}{5} \Longrightarrow x = 70 + 2.3(5) = 81.5 \text{ years}$$

Remark. You can always compute the z-score of a point given μ and σ.

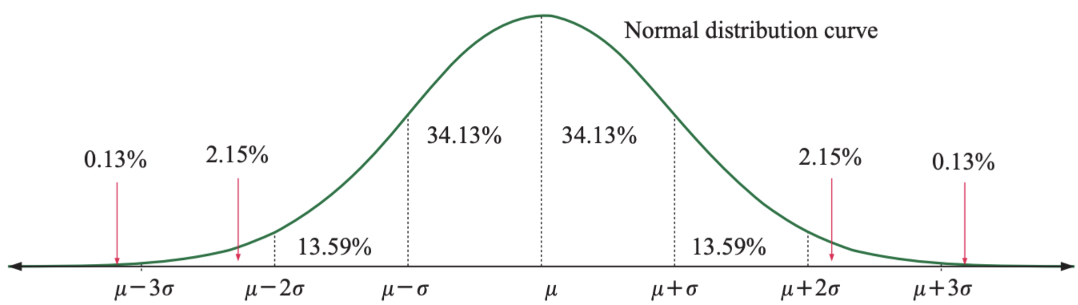
The only reason why we care about these z-scores is that they give us percentile interpretations when our distribution is *normal*.

3. Normal Distributions

When we are given a **normal distribution**, we write $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation. We know that X is going to be normal distribution if the associated PDF is given by the function

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

This PDF will look like the bell-curve but also satisfies our 68-95-99.7 empirical rule.



Note that

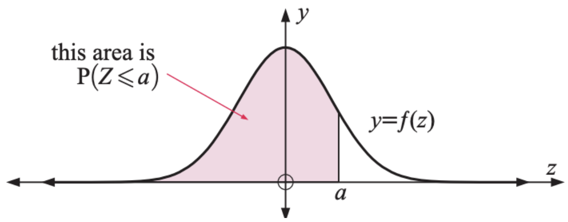
- $\approx 68.26\%$ of the data lives between $\mu - \sigma$ and $\mu + \sigma$
- $\approx 95.44\%$ of the data lives between $\mu - 2\sigma$ and $\mu + 2\sigma$
- $\approx 99.74\%$ of the data lives between $\mu - 3\sigma$ and $\mu + 3\sigma$

When the distribution is normal, we get an easy relationship between **percentiles** and **z-score**.

Example 1. If X is a normal distribution with $\mu = 62$ and $\sigma = 7$, let's compute

(1) percentile of $X = 69$
(2) Find the probability $Prob(58.5 \leq X \leq 71.8)$.

The tool to use is the z-score and recall that percentiles start from the left:



So compute the z-score of $X = 69$

$$z_{69} = \frac{69-\mu}{\sigma} = \frac{69-62}{7} = 1$$

so we know that $X - 69$ is exactly 1 standard deviation above the mean. Using the z-score table, we get that $X = 69$ is in the 84.13rd percentile.

For (2), we find the z-scores of 58.5 and 71.8, respectively:

$$z_{58.5} = \frac{58.5-62}{7} = -\frac{1}{2}$$
$$z_{71.8} = \frac{71.8-62}{7} = 1.4$$

so locating these onto the z-score table, we find that $z_{71.7}$ is in the 91.92nd percentile while $z_{71.8}$ is in the 30.85 percentile. So we conclude that the probability $Prob(58.5 \leq X \leq 71.8)$ if $X \sim N(62, 7)$ is going to be $0.9192 - 0.3085 = 0.611$. So there's 61.07%.

3.1. **The 3-Step Plan.** To find probabilities for a normally distributed variable,

(1) Convert the X-values into z-scores
(2) Sketch a standard normal curve and shade in the required region asked by the question
(3) Use a z-score table (or calculator) to find probabilities