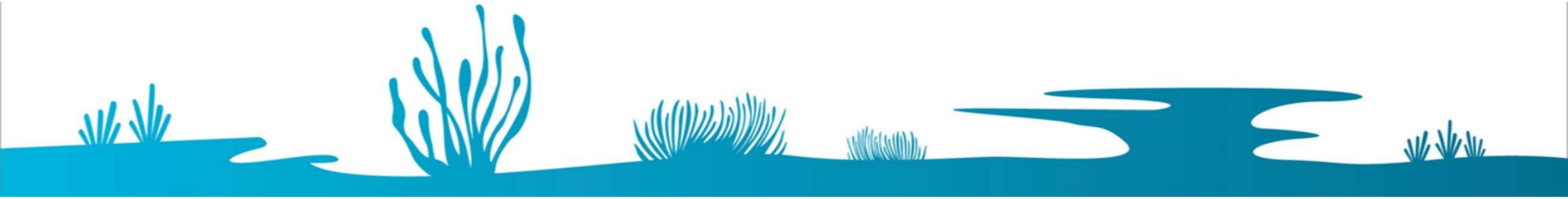
Climate and Oceans Support
Program in the Pacific

ACCESS-S Workshop

MODULE: Statistics – quantiles and probabilities



Topics in this module

- Quantiles and distributions
- Quantiles and probability

Expected learning outcomes

- Understanding distributions and their relationship to probability

Climate and Oceans Support Program in the Pacific

Standard deviation and variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

How spread out is the dataset from the mean?
Standard deviation and variance can tell us

- S = **Sample Standard Deviation**
- n = Sample size
- $x_i - \bar{x}$ = An observed **anomaly**

1. Each anomaly is squared

2. We calculate the sum of all the squared anomalies

3. This sum is divided by a number one less than the sample size, i.e. n – 1

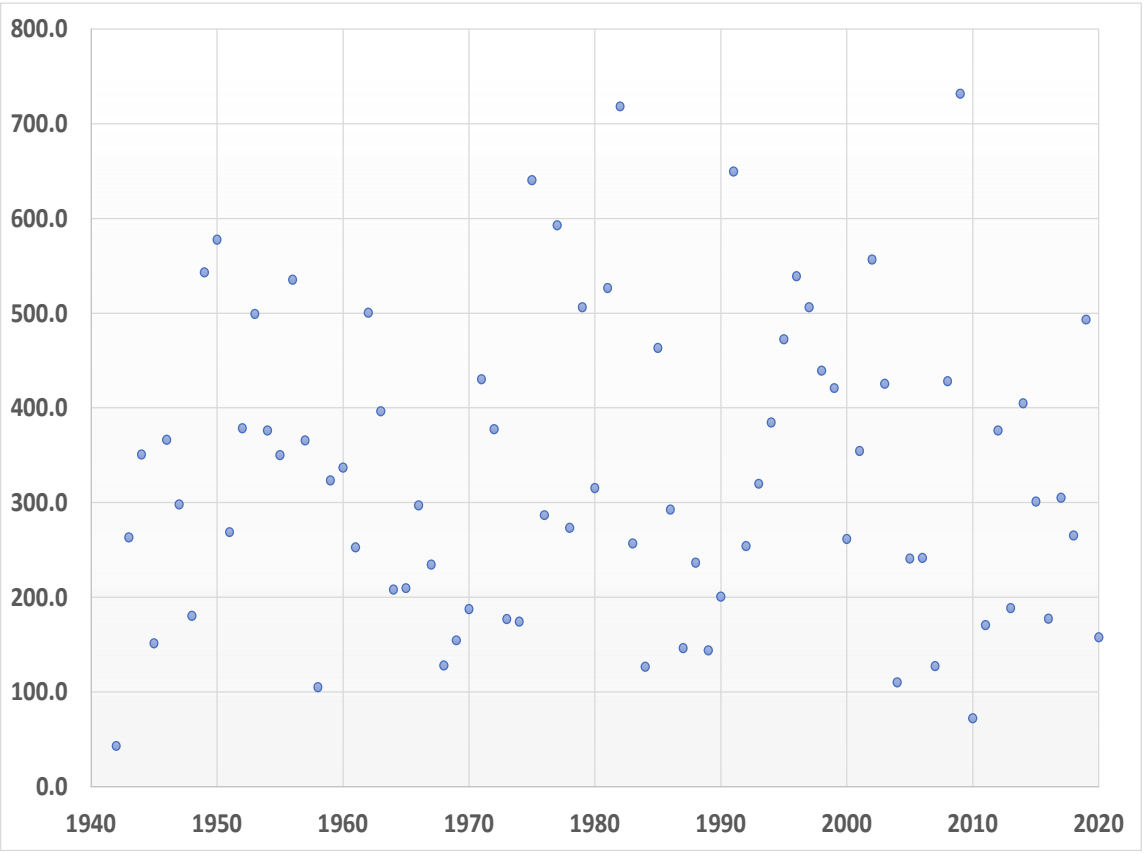
4. Finally, calculate the square root of the value calculated at this point

- This is the Standard deviation
- The Variance = (Std Dev)^2
- Or Std Dev = $\sqrt{Var}$

Climate and Oceans Support Program in the Pacific

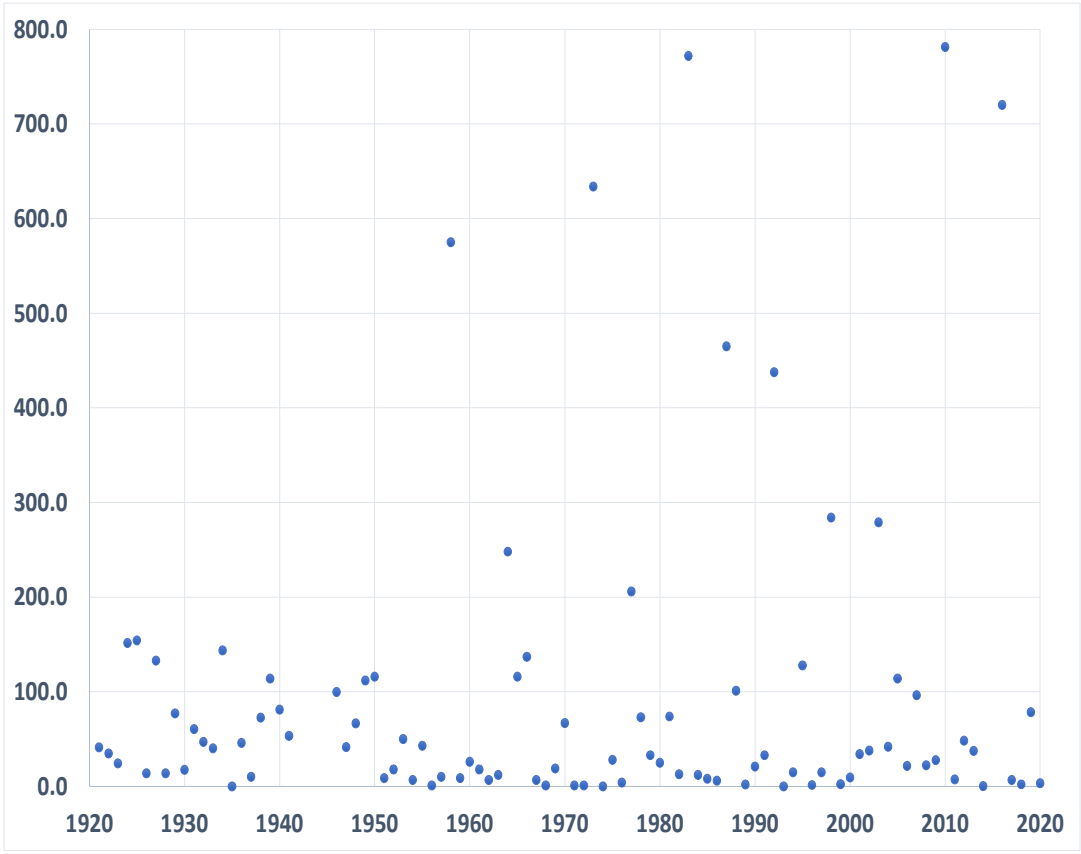
Standard deviation and variance

Suva, Fiji. January Rainfall



Mean 331mm, S.D. 158mm

Kiritimati, Kiribati. January Rainfall



Mean 95mm, S.D. 167mm

Climate and Oceans Support Program in the Pacific

Quantiles and distributions

In climatology observations occur in **time**, or **chronologically**

A **distribution** is the data ordered lowest to highest, or **numerically**

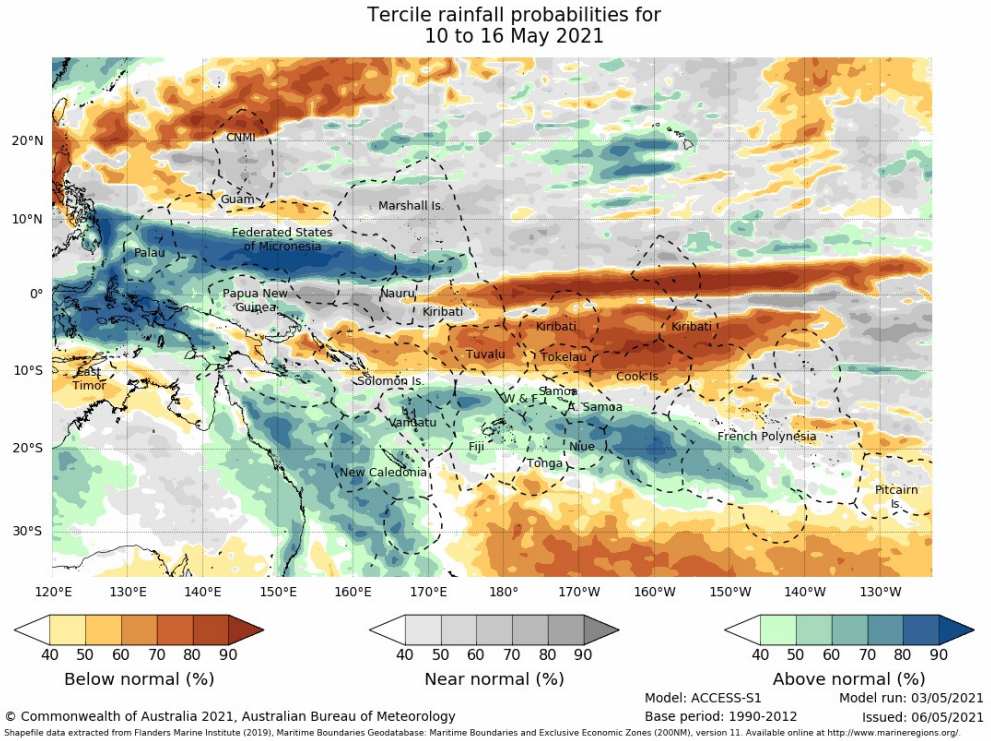
If we split the **distribution in two** we have calculated the **median**.

If we split the **distribution into three** we have calculated **terciles**.

Terciles are an ACCESS-S output

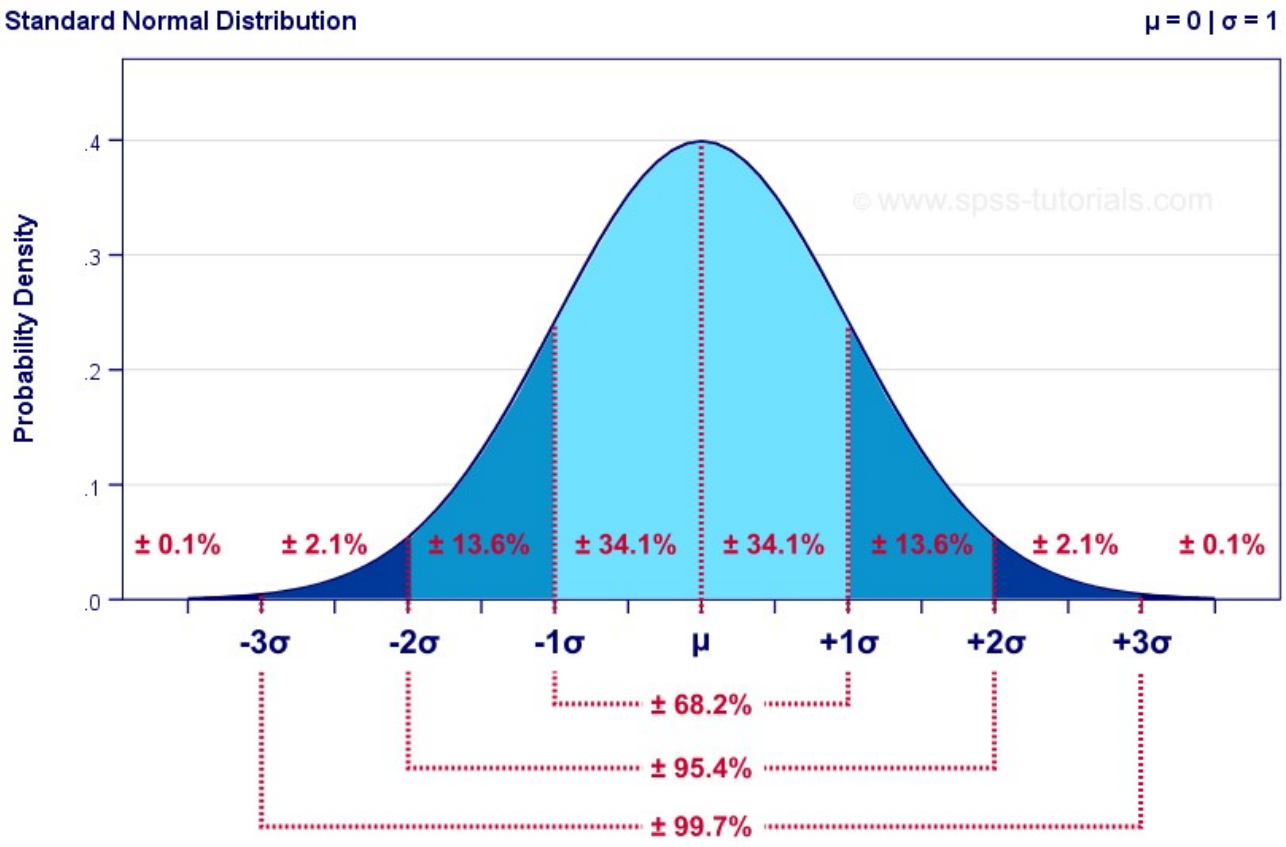
The **median and terciles** are both common "**quantiles**", which describes splitting the distribution

Common quantiles:
Median (halves): or 50th percentile
Terciles (thirds): 33.3rd and 66.7th percentiles
Quartiles (quarters): 25th, 50th, and 75th percentiles
Deciles (tenths): 10th, 20th, … , 80th, 90th percentiles



Climate and Oceans Support Program in the Pacific

The Normal Distribution



If the data is symmetrical it fits into the Normal Distribution

- μ = the **mean** (zero in standard normal)
- σ = the **standard deviation** (1 in standard normal)
- Symmetrical **Bell-Shaped** curve
- **Mean = Median** in this distribution
- Probability Density (Y Axis) is equivalent to frequency (number of observations)
- 68.2% of observations are within 1 σ of the mean/median
- 95.4% of observations are within 2 σ of the mean/median
- 99.7% of observations are within 3 σ of the mean/median
- In the **Pacific**, temperatures are more likely to be normally distributed than rainfall

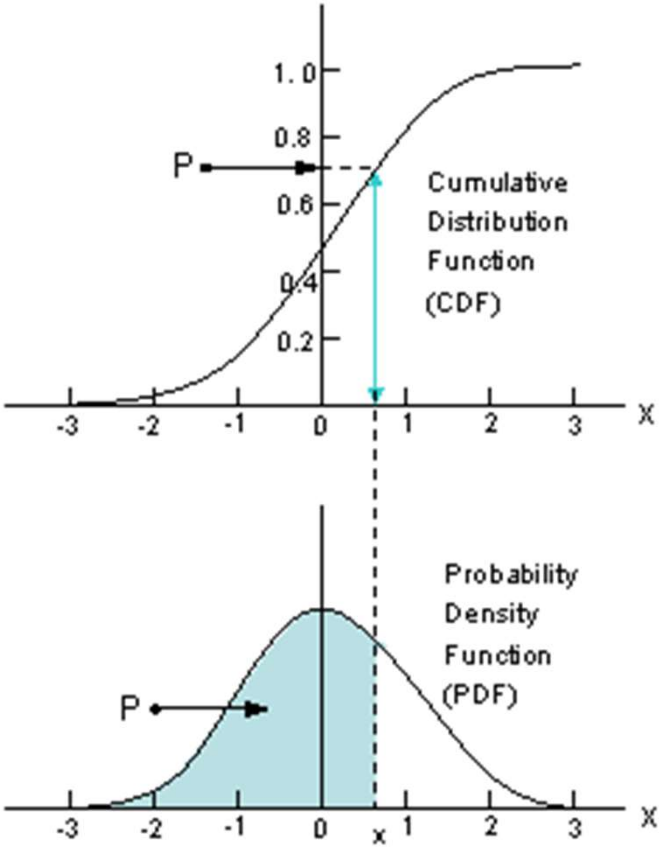
Climate and Oceans Support Program in the Pacific

Quantiles and Probabilities

- By definition, **quantiles split a distribution into two parts**, with one fraction or percentage of the observations being below the quantile and the remaining fraction lying above
- We've already seen the Median (50th percentile) splits the distribution in half
- Other Examples:
  - the 10th percentile divides the lowest 10% of observations from the remaining 90%
  - the 67th percentile divides the lowest two-thirds of observations from the remaining 33%
  - the 90th percentile divides the lowest 90% of observations from the remaining 10%

Climate and Oceans Support Program in the Pacific

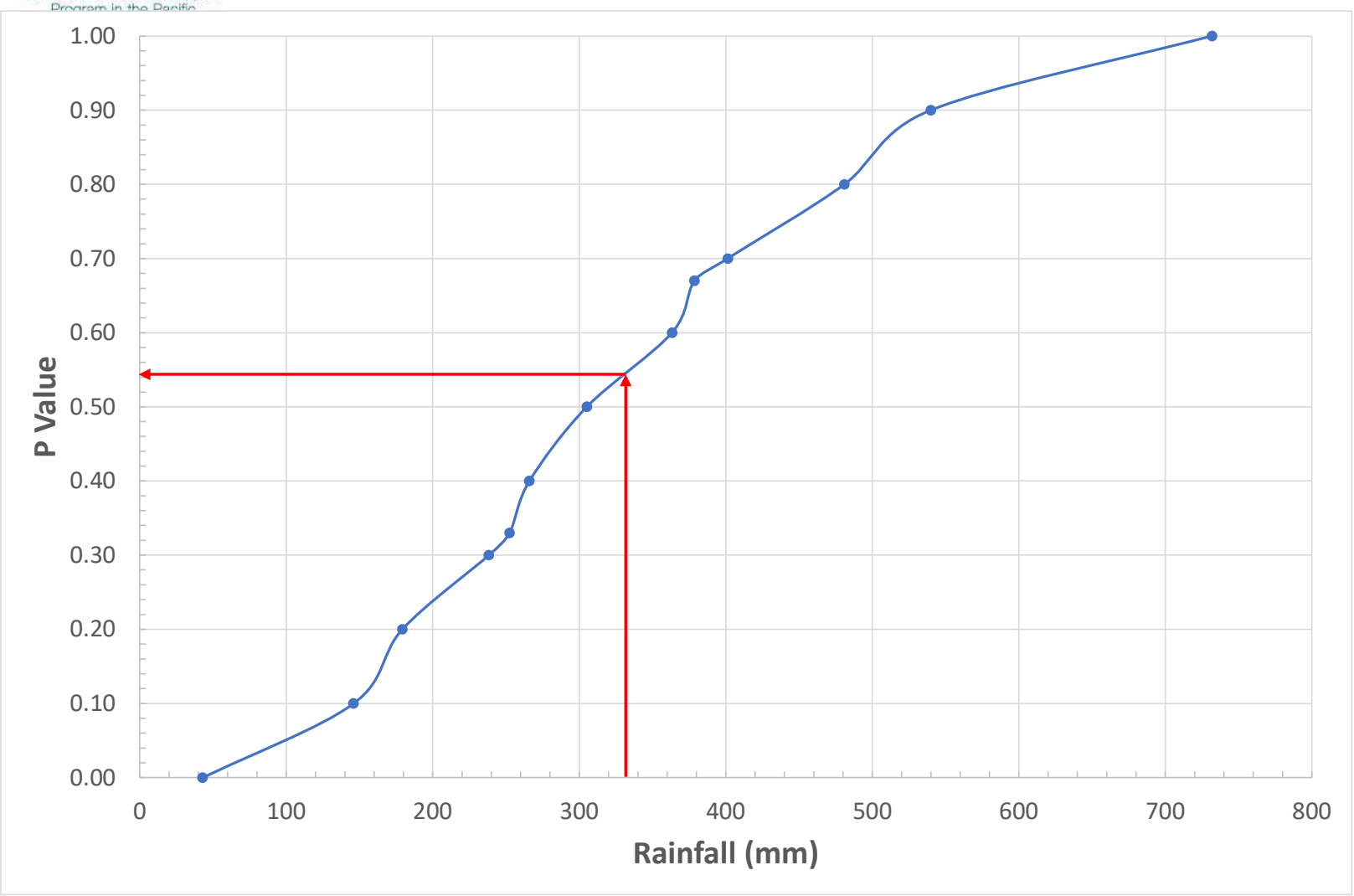
The Cumulative Distribution



Relations Between Two Different Typical Representations of a Population

- Lower graph = normal distribution
- Shaded area on lower graph (for a value x) represents the fraction or proportion of values ≤ x. This is called **P**
- **P** ranges from **0 to 1**
- If we plot values of P for increasing values of x, we create the upper graph
- For given values of P, we can find a value of x, e.g.
  P = 0.50, x = median
  P = 0.33, x = 33rd percentile (top of Tercile 1)

Cumulative Distribution: Suva, Fiji, January Rainfall



The Cumulative Distribution

- Each dot is a different percentile
- Dots included: Lowest, Highest, All Deciles, Terciles
- **January mean (331mm):** Follow red arrow up from 331 to the blue graph, then across to the Y axis to read the P value (=0.54 in this case)
- Mean ≠ Median in this example
- Therefore, Suva's January rainfall is **not** normally distributed

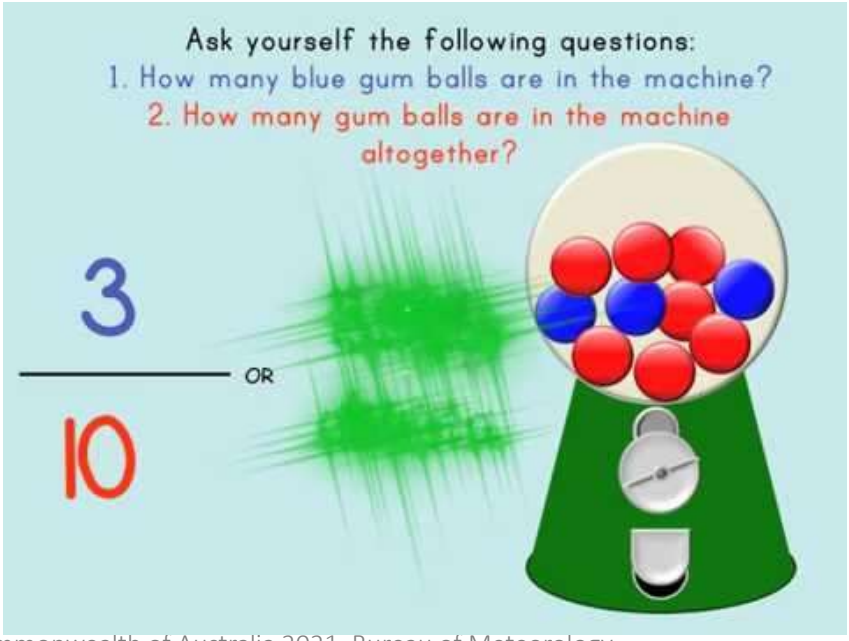
Climate and Oceans Support Program in the Pacific

Quantiles and Probabilities

We can Generalise:

For the Nth Percentile, N% of observations are below and (100-N)% are above

Example: If we picked any year at random, the chance or **probability** that we select a year in which Suva's January rainfall was **above** the 75th percentile is 25% (100-75).

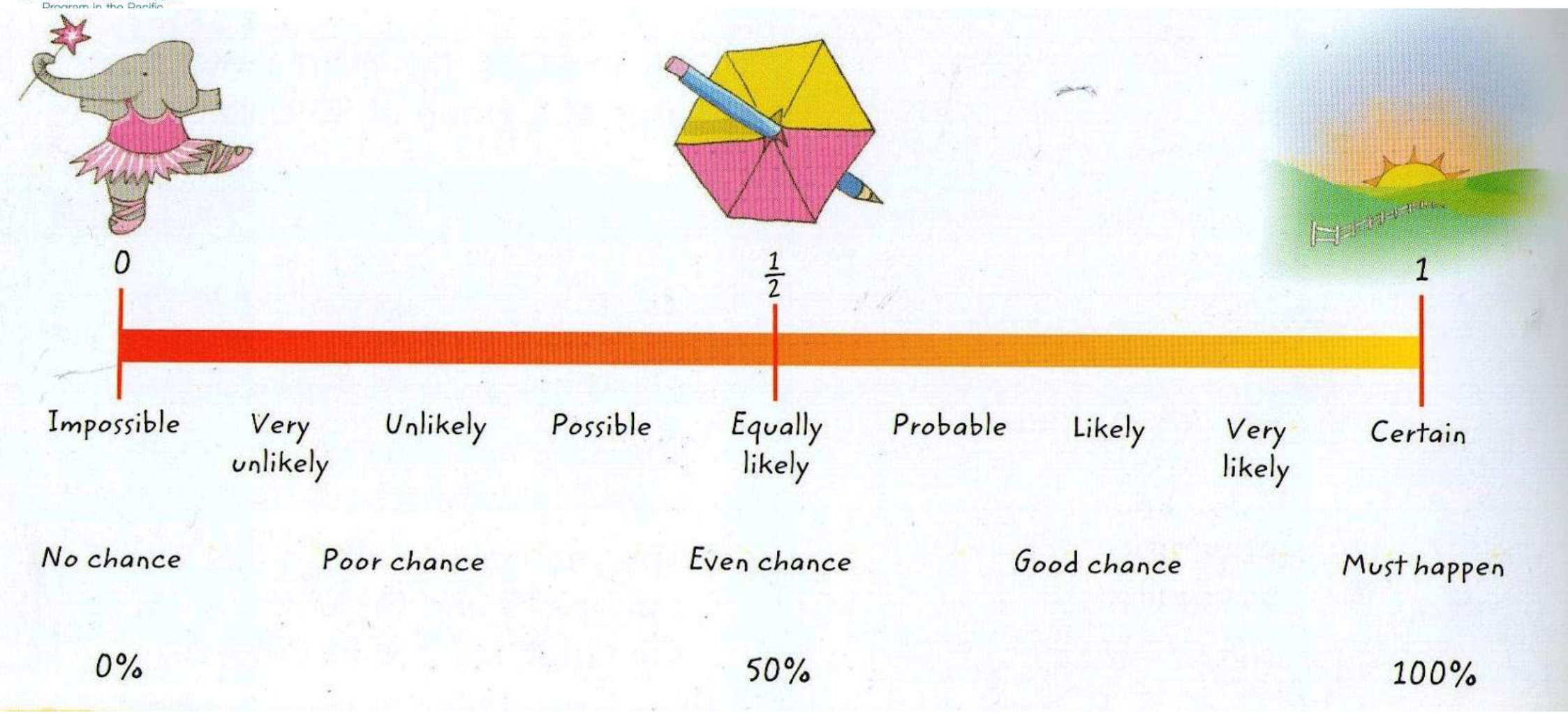


Consider the gum ball dispenser

Let the blue balls represent the lowest three deciles (30th percentile or lower) of observations (e.g. rain, temperature, thunder days, etc.), and the red balls represent the highest seven deciles.

The probability of getting a blue ball is 30%.

Language and Probabilities



Quantiles and Distributions summary

A **distribution** is the data ordered lowest to highest, or **numerically**

A **quantile splits a distribution into two parts**, with one fraction or percentage of the observations being below the quantile and the remaining fraction lying above

The **median** splits the distribution in half

Terciles split the distribution into three parts

Quantiles of a **distribution** can give us the **probability**

There is language associated with probability, to help users understand how likely something will happen

Discussion questions

What are some common quantiles you have used or seen?

How do you use language to describe different probabilities?