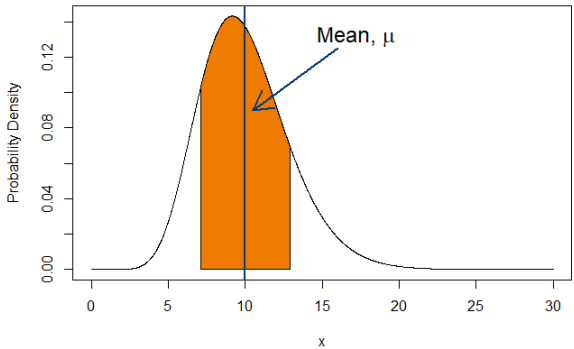
Numerical Summaries of Probability Distributions



Outline

Learning Objectives

By the end of this video, we hope that you will be able to:

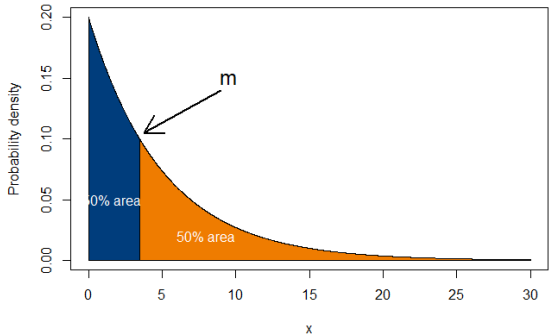
- Understand the numerical summaries in probability distributions.
- Understand the relationship between observed data and probability distributions.

Median of Probability Distributions

Median of a Probability Distribution

Median of a set of numbers

The *middle number* of the ordered set of those numbers - half the dataset is below that number, and half are above it.



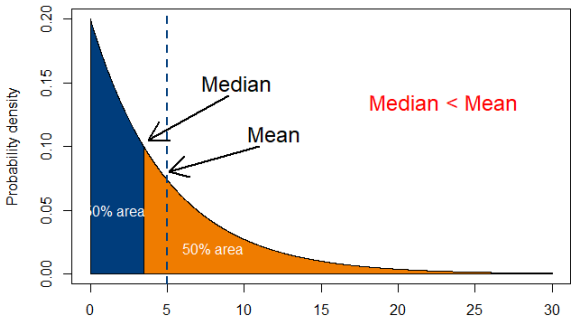
- "m" is known as the median of the probability distribution

Mean of Probability Distributions

Mean of a Probability Distribution

Mean of a set of numbers

- The *average* of that set of numbers.
- In probability distribution, it is the average of all possible values, *weighted by the probabilities*, that the random variable could take on.



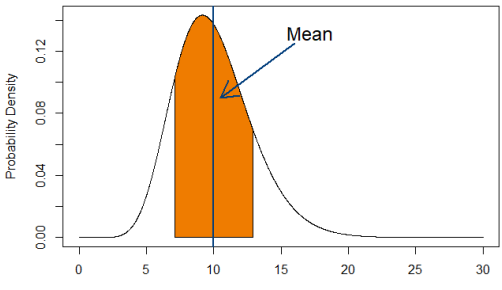
- **Mean** and **median** of distributions provide indications of the "middle" of the distribution, and are known as measures of central tendency.

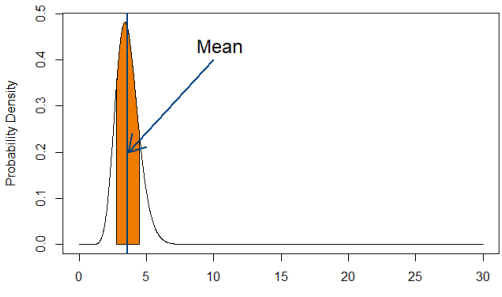
Variance of Probability Distributions

Variance of a Probability Distribution

Variance of a set of numbers

- An indication of *how spread apart* the data was. The larger the sample variance was, the more values we had that were far away from the sample mean.
- The variance of a probability distribution is an indication of how far from the mean values could possible be, when we make observations of things from that distribution.





Quantiles of Probability Distributions

Quantiles

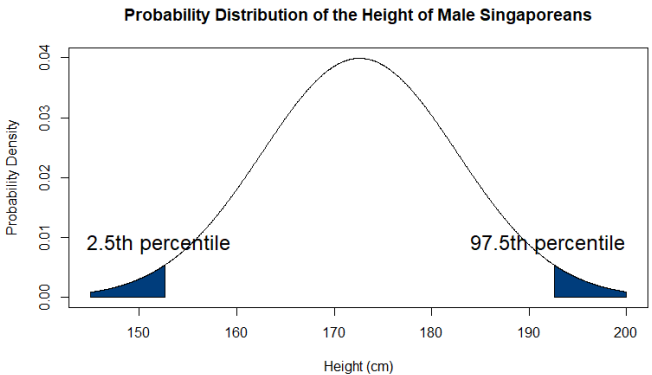
Quantile

In statistics, quantile specifies the value of the random variable such that th e probability of the variable being *less than* or *equal to* that value equals the given probability.

- Quantile helps us to define a particular part of a probability distribution by determining how many values in a distribution are *above* or *below* a certain limit.
- We use **percentile** more commonly in this module.

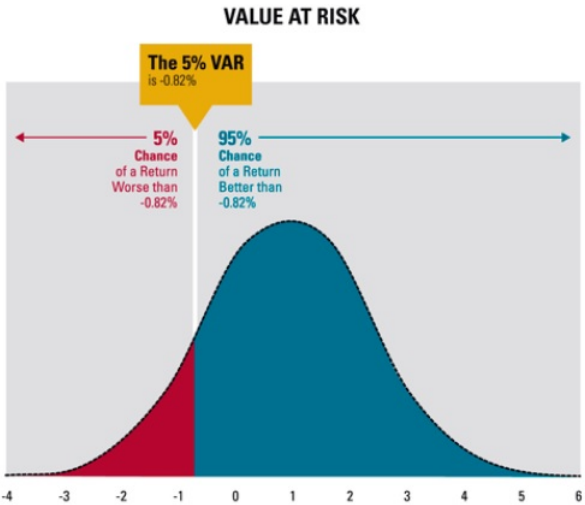
Quantiles

cont'd



- 2.5% of the male Singaporean population will have a height of *more than* 193 cm.
- 2.5% of the male Singaporean population will have a height of *less than* 152 cm.

Application of Quantiles



Value at risk (VaR) (Abbasi and Guillen, 2013)

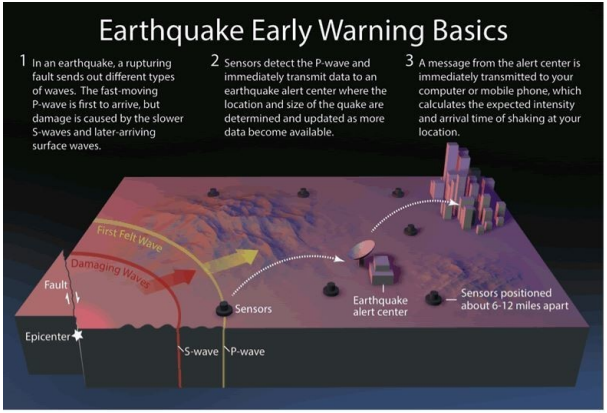
Used by investment banks or insurance firm to determine the extent and probabilities of potential financial losses in their institutional portfolios.

Application of Quantiles

cont'd

Earthquake Early Warning System (Pisarenko and Rodkin, 2013)

Seismologists use quantile approach to assess any risk of seismic activities. Helps to predict upcoming rare earthquakes.



Observed Data and Properties of Distributions

Relationship between Observed Data and Distributions

- When we record data in a single column in our dataset, we are often assuming that they come from the same distribution.
- Denote the relationship with:

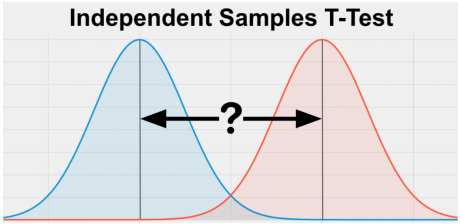
$$X_1, X_2, ..., X_n \sim f(x)$$

  - read as the X_i's are *independent and identically distributed* random variables.
- If the number of observations we have is large enough, then the numerical summaries computed from the data are good estimates of the distribution quantities.
  - Mean of the n values will be a good estimate of the distribution mean.
  - Variance of the n numbers will be a good estimate of the variance of f.
- We use observed quantities to **estimate** the properties of the distribution from which they came.
  - The more observations we have, the better the estimate will be.

Relationship between Observed Data and Distributions

cont'd

- When performing **hypothesis testing** or **confidence intervals**, we are comparing the properties of the distributions from which the data arose.



Observations:

First group	Second group
$X_1, X_2, ..., X_n \sim f_1(x)$	$Y_1, Y_2, ..., Y_n \sim f_2(x)$

- When performing two sample t-test, we are essentially checking if the mean of distribution f_1 is the same as the mean of distribution f_2.

Summary

Learning Outcomes

- Able to understand the numerical summaries of probability distributions: Median, Mean, Variance, Quantile.
- Able to understand the relationship between observed data and the properties of probability distributions.

References

Abbasi, B. and Guillen, M. (2013). Bootstrap control charts in monitoring value at risk in insurance. *Expert Systems with Applications*, 40(15):6125–6135.

Diez, D. M., Barr, C. D., and Mine, C.-R. (2019). page 115–123. OpenIntro, Inc., 4th edition.

Pisarenko, V. and Rodkin, M. (2013). The new quantile approach: Application to the seismic risk assessment. *Natural Disasters: Prevention, Risk Factors and Management*, pages 141–174.

Ross, S. M. (2020). *A first course in probability*. Pearson Education Limited.