



BITS Pilani
DIGITAL
Excellence made yours








BITS Pilani
DIGITAL
Excellence made yours

Probability and Statistics

Week 3 Live Session

Dr. Pushpendra Gupta

Today's Focus

-  Recap
-  Chebyshev's Inequality in Practice
-  Visualization Techniques
-  Comparing quartile methods
-  Hands-on Visualization using Google Colab

Week 1 Recap



Central Tendency



Variability Measures



Data Visualization



Distribution Analysis

Week 2 Recap



Quartile Deviation and IQR



5 Point Summary



Box Plot



Infer from Statistical Summary

Week 3 Recap



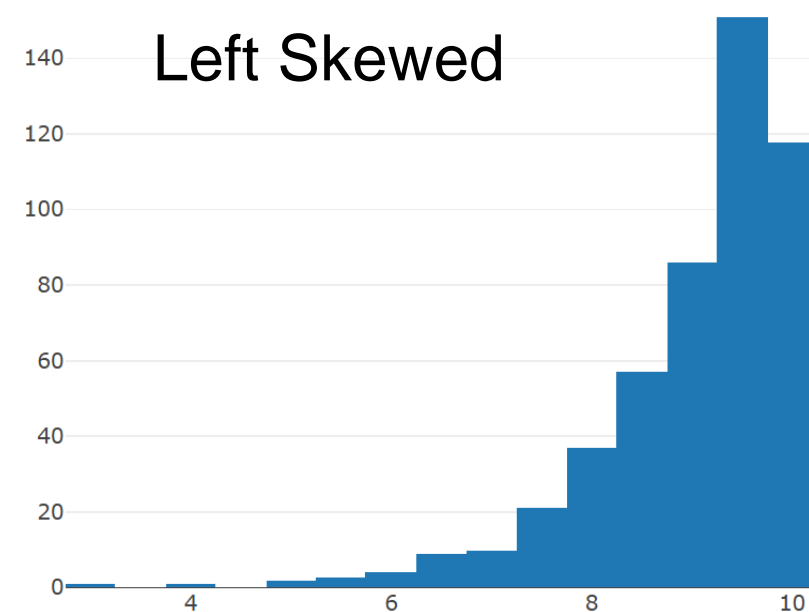
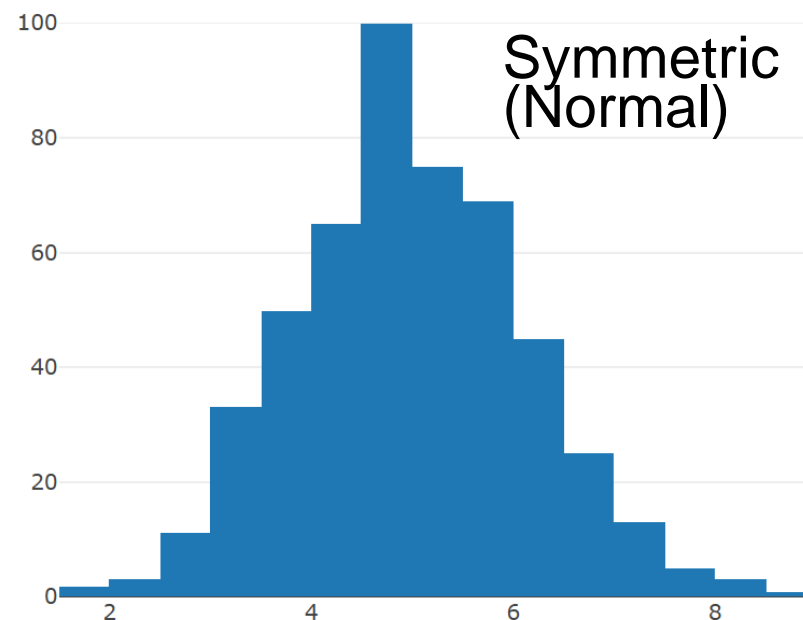
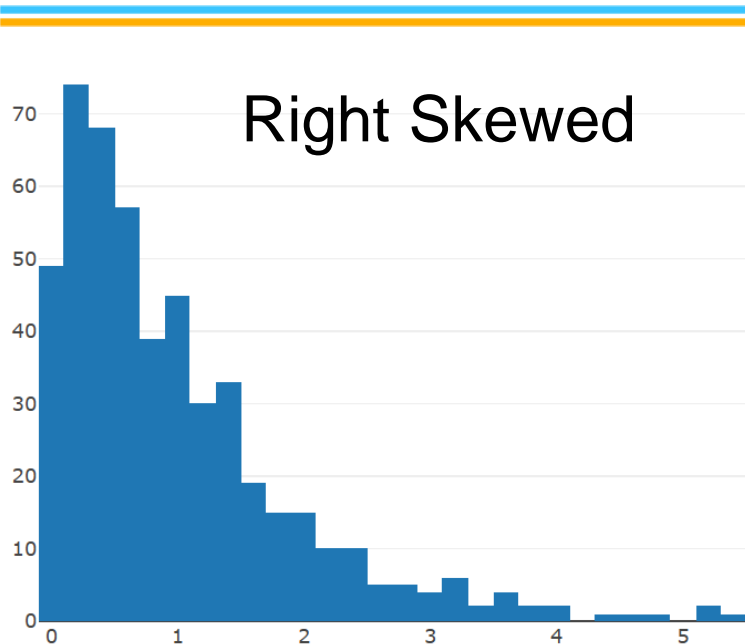
Element of Probability

Combinations

Finding probabilities



Chebyshev's Inequality in Practice



For ANY dataset, at least $\left(1 - \frac{1}{k^2}\right)$ of data lies within k standard deviations

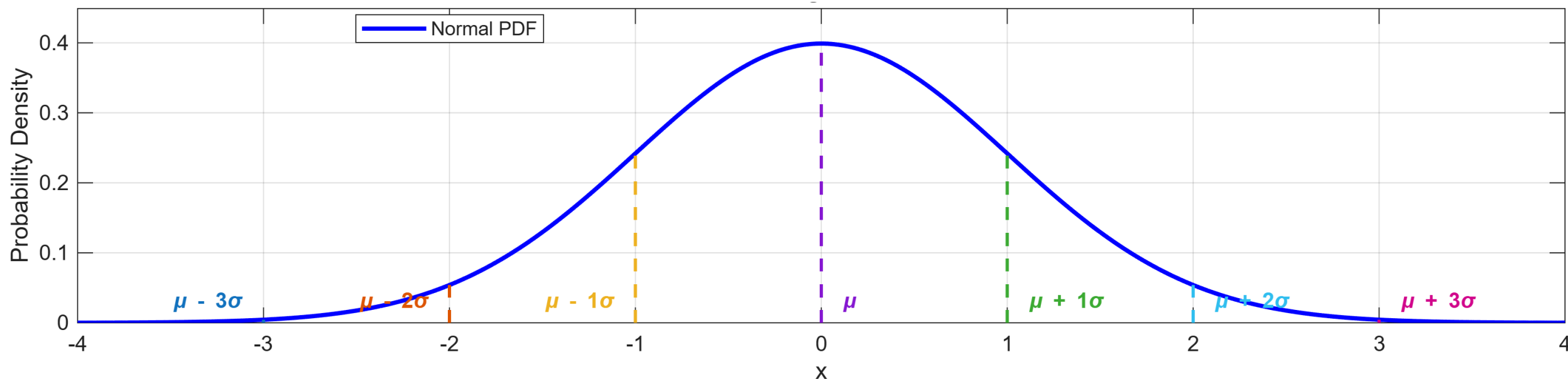
at most $\frac{1}{k^2}$ of the data will be outside k standard deviations from the mean.



Chebyshev's Inequality in Practice



BITS Pilani
DIGITAL
Excellence made yours



| k | At most outside $k\sigma$ | At least within $k\sigma$ | Normal comparison |
|-----|---------------------------|---------------------------|-------------------|
| 1.5 | 44.4% | 55.6% | 86.6% within |
| 2 | 25% | 75% | 95% within |
| 2.5 | 16% | 84% | 98.8% within |
| 3 | 11.1% | 88.9% | 99.7% within |



Two Forms of Chebyshev's Inequality

Form 1: Upper Bound (for outliers)

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Interpretation: At most $\frac{1}{k^2}$ of the data will be outside k standard deviations from the mean.

Form 2: Lower Bound (for central data)

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Interpretation: At least $\left(1 - \frac{1}{k^2}\right)$ of the data will be within k standard deviations from the mean.



Chebyshev's Inequality in Practice

A payment system processes transactions with mean Rs. 500 and standard deviation Rs. 150. Without knowing the distribution, what can we say about transactions outside normal ranges?

using Chebyshev with $k = 2$:

- Range: $\mu \pm 2\sigma = 500 \pm 300 = [200, 800]$
- Chebyshev guarantees: At most 25% of transactions fall outside $[200, 800]$
- Equivalently: At least 75% of transactions are between 200 and 800

This 25% is an upper bound on outliers, not the probability that any specific flagged transaction is fraudulent!

$$P(\text{transaction outside } [200, 800]) \leq 0.25$$



Chebyshev's Inequality in Practice



BITS Pilani
DIGITAL
Excellence made yours

A manufacturing process produces widgets with mean weight 100g and standard deviation 5g. What percentage of widgets must weigh between 85g and 115g?

This range is $\mu \pm 3\sigma$ ($k = 3$) By Chebyshev's inequality:

At least $1 - 1/9 = 88.9\%$ of widgets fall in this range.



Chebyshev's Inequality in Practice



BITS Pilani
DIGITAL
Excellence made yours

| Serial No. | Student Transactions (Rs.) | Business Transactions (Rs.) |
|------------|----------------------------|-----------------------------|
| 1 | 210 | 11,200 |
| 2 | 280 | 8850 |
| 3 | 190 | 33,100 |
| ... | ... | ... |
| 100 | 320 | 9950 |

Assume Student data has:

Mean (μ) = Rs. 250, Standard Deviation (σ) = Rs. 50

$$\mu \pm 2.5\sigma = 250 \pm 2.5 \times 50 = [Rs. 125, Rs. 375]$$



Chebyshev's Inequality in Practice



BITS Pilani
DIGITAL
Excellence made yours

Assume Student data has:

Mean (μ) = Rs. 250, Standard Deviation (σ) = Rs. 50

$$\mu \pm 2.5\sigma = 250 \pm 2.5 \times 50 = [Rs. 125, Rs. 375]$$

84% of transactions will be within [Rs. 125, Rs. 375]. (use $1 - 1/k^2$)
At most 16% (≤ 16 transactions out of 100) fall outside this range.

But What If Data Is Normal?

Actual coverage for $\mu \pm 2.5\sigma$ jumps to $\sim 98.76\%$

Only $\sim 1.24\%$ (≈ 1 transaction) would be outside [Rs. 125, Rs. 375].



Chebyshev's Inequality in Practice

Consider 100 coin flips with $p = 0.5$. What's the probability of getting 70 or more heads?

$$n = 100, p = 0.5,$$

$$\text{Mean} = np = 50, \text{ Standard deviation} = \sqrt{np(1-p)} = 5$$

70 heads is 4 standard deviations from the mean.

$$\text{Chebyshev's Bound } P(|X - 50| \geq 4 \times 5) \leq 1/4^2 = 0.0625$$

$$\text{Meaning } P(X \leq 30 \text{ or } X \geq 70) \leq 6.25\%$$



Chebyshev's Inequality in Practice

Consider 100 coin flips with $p = 0.5$. What's the probability of getting 70 or more heads?

However Exact probability can only be found using Binomial

$$P(X \geq 70) = \frac{100!}{70! 30!} \times 0.5^{70} \times (1 - 0.5)^{30} \approx 0.000003$$

Chebyshev gives a conservative upper bound (6.25%) while the actual probability is much smaller (0.0003%).



Frequency Tables

A frequency table organizes data by showing how often each value or category appears.

(Monthly Sales Data). Consider monthly sales (in thousands) from a small retail store over 12 months: 45, 52, 48, 55, 62, 58, 51, 49, 56, 60, 53, 57

| Sales Range | Frequency | Relative Frequency |
|-------------|-----------|--------------------|
| 45-49 | 3 | 0.25 |
| 50-54 | 3 | 0.25 |
| 55-59 | 4 | 0.33 |
| 60-64 | 2 | 0.17 |

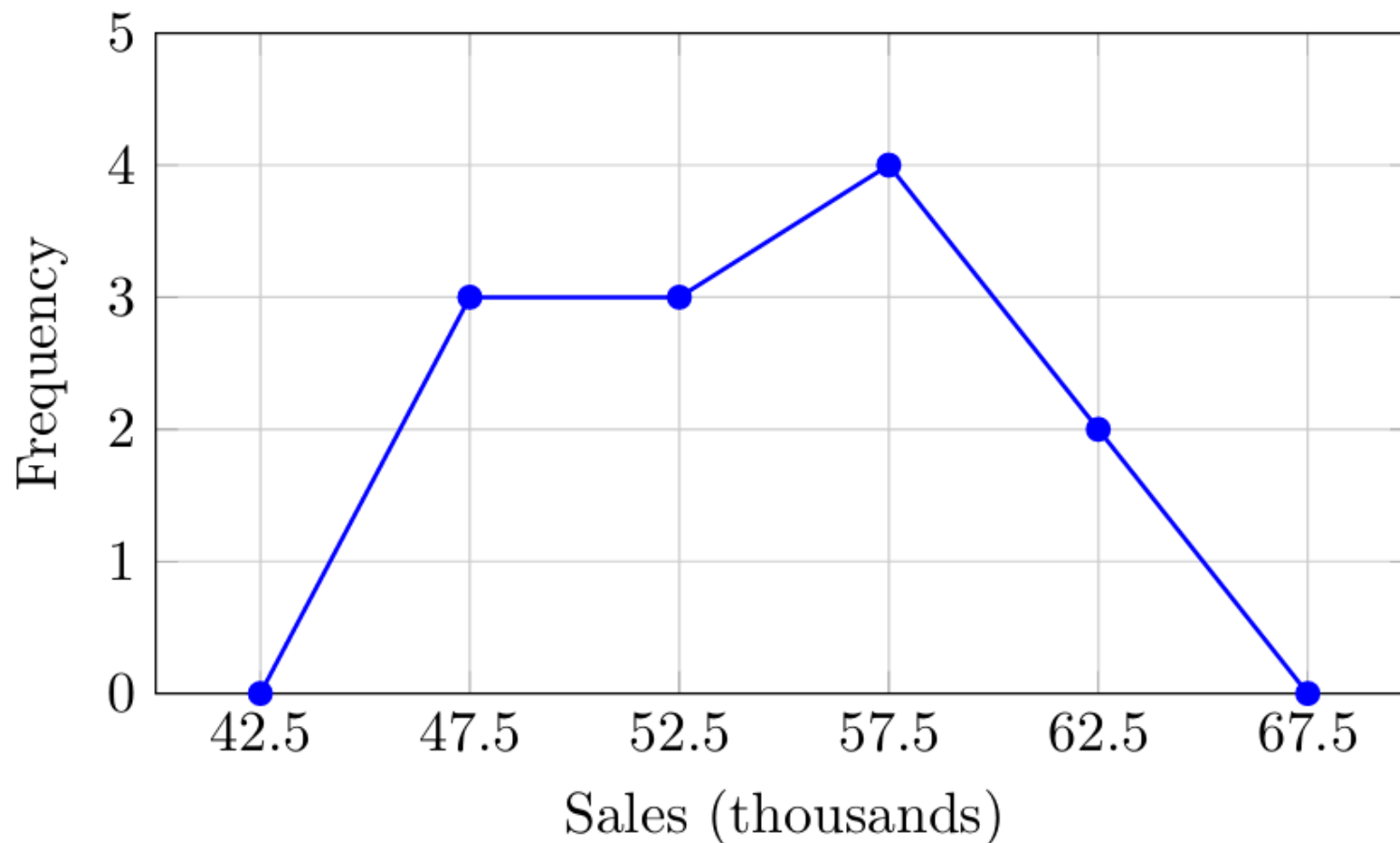


Frequency Polygon - Continuous Class



Continuous Classes have no gaps between successive intervals. The upper limit of one class equals the lower limit of the next class.

| Sales Range | Frequency | Midpoint Calculation |
|-------------|-----------|----------------------|
| 45-50 | 3 | $(45+50)/2=47.5$ |
| 50-55 | 3 | 52.5 |
| 55-60 | 4 | 57.5 |
| 60-65 | 2 | 62.5 |





Frequency Polygon- Discontinuous Class



BITS Pilani
DIGITAL
Excellence made yours

Discontinuous classes have gaps between successive intervals.

e.g., classes like 20-24, 25-29, 30-34 have a gap of 1 unit (24 to 25)

The gap creates ambiguity. Where would a value of 24.5 fall? To resolve this, we create class boundaries that make the intervals continuous.

| Sales Range | Frequency | Lower Boundary | Upper Boundary | Midpoint |
|-------------|-----------|-------------------|-------------------|--------------------------|
| 20-24 | 4 | $20 - 0.5 = 19.5$ | $24 + 0.5 = 24.5$ | $(19.5 + 24.5)/2 = 22.0$ |
| 25-29 | 7 | $25 - 0.5 = 24.5$ | $29 + 0.5 = 29.5$ | $(24.5 + 29.5)/2 = 27.0$ |
| 30-34 | 10 | $30 - 0.5 = 29.5$ | $34 + 0.5 = 34.5$ | $(29.5 + 34.5)/2 = 32.0$ |
| 35-39 | 6 | $35 - 0.5 = 34.5$ | $39 + 0.5 = 39.5$ | $(34.5 + 39.5)/2 = 37.0$ |
| 40-44 | 3 | $40 - 0.5 = 39.5$ | $44 + 0.5 = 44.5$ | $(39.5 + 44.5)/2 = 42.0$ |

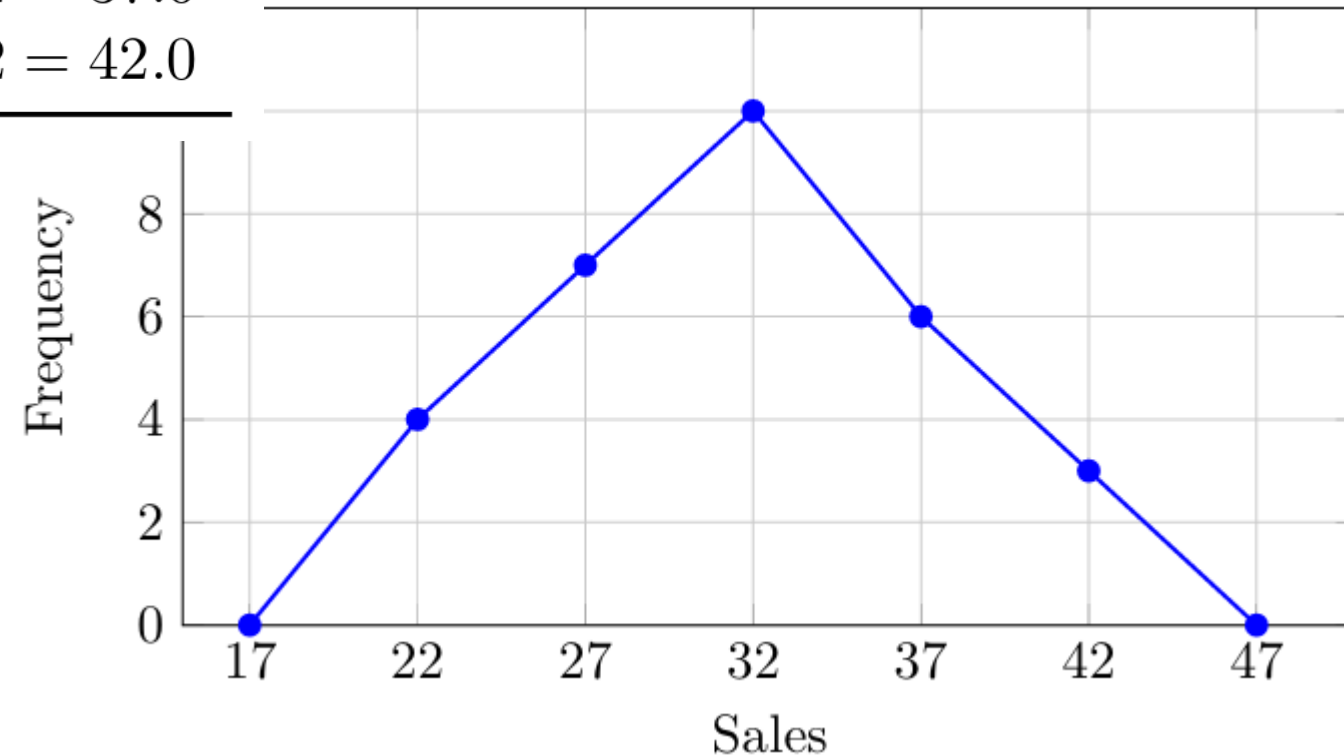


Frequency Polygon- Discontinuous Class



BITS Pilani
DIGITAL
Excellence made yours

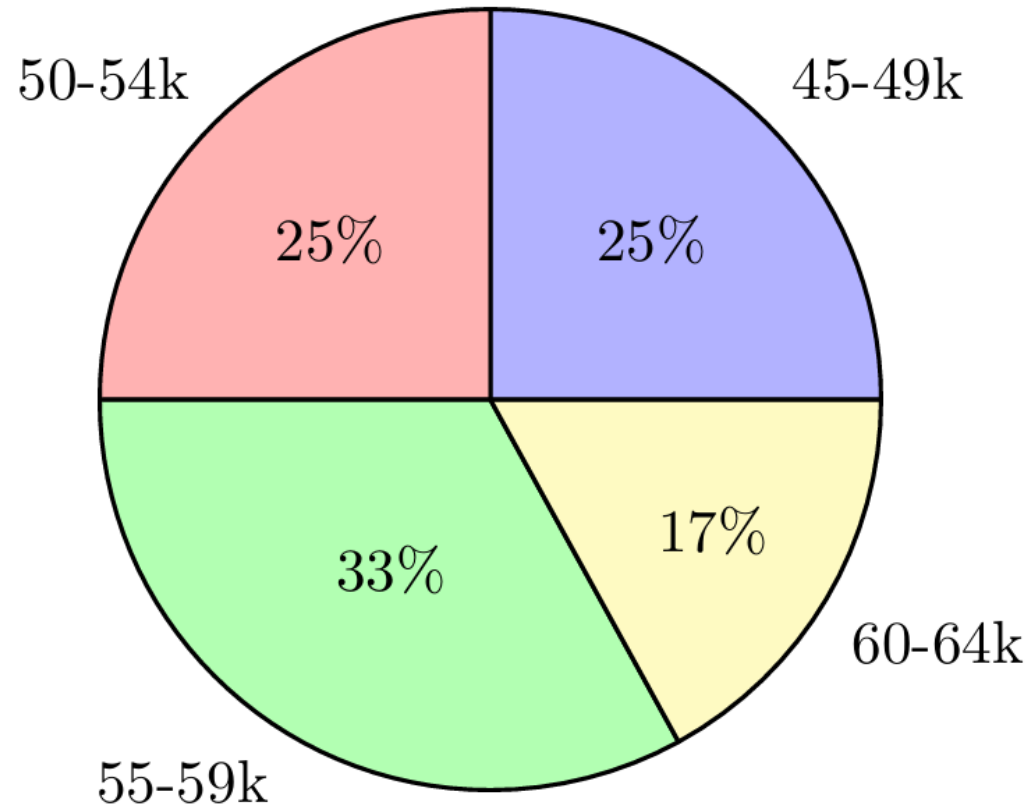
| Sales Range | Frequency | Midpoint |
|-------------|-----------|--------------------------|
| 20-24 | 4 | $(19.5 + 24.5)/2 = 22.0$ |
| 25-29 | 7 | $(24.5 + 29.5)/2 = 27.0$ |
| 30-34 | 10 | $(29.5 + 34.5)/2 = 32.0$ |
| 35-39 | 6 | $(34.5 + 39.5)/2 = 37.0$ |
| 40-44 | 3 | $(39.5 + 44.5)/2 = 42.0$ |





Pie Charts

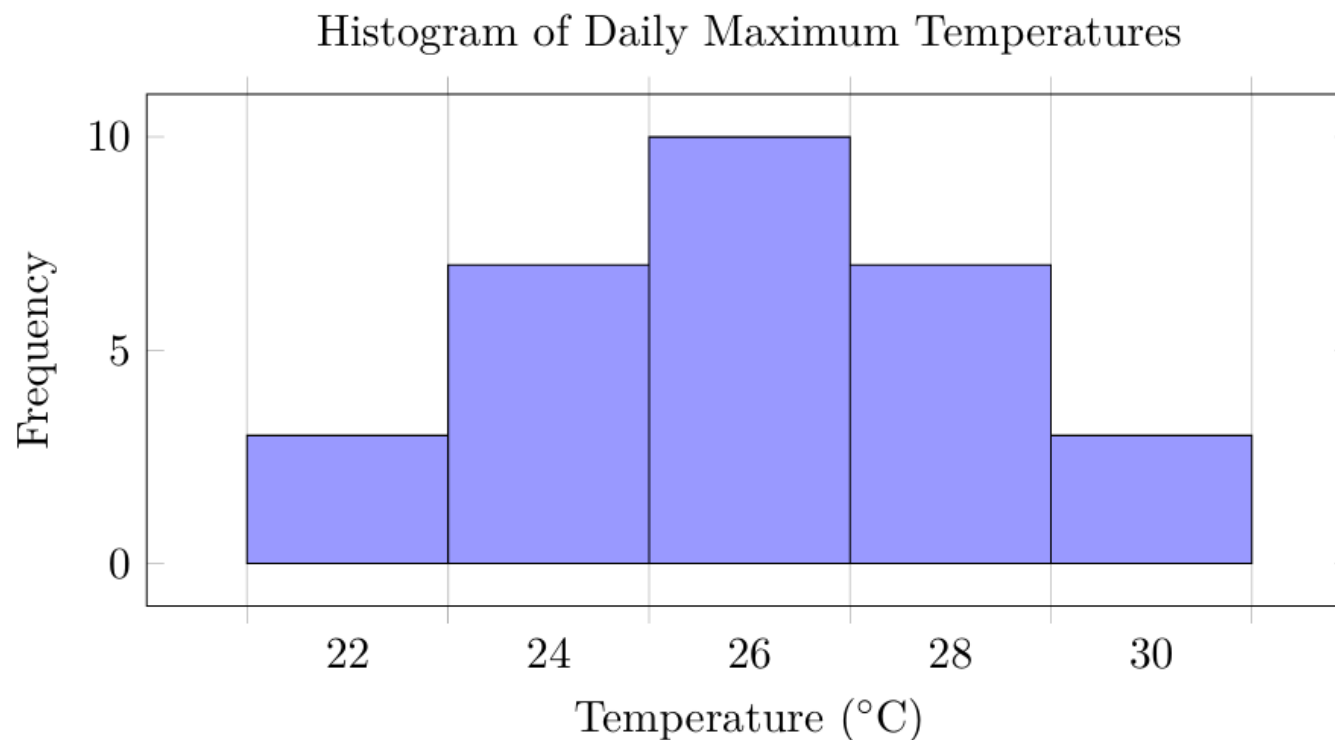
Pie charts display proportions effectively for categorical data.





Histograms

Histograms group continuous data into bins and display frequencies as rectangular bars. (Temperature Data). Daily maximum temperatures ($^{\circ}\text{C}$) for 30 days: 22, 24, 23, 26, 28, 25, 27, 29, 24, 26, 31, 30, 28, 25, 27, 23, 24, 26, 28, 29, 27, 25, 24, 26, 30, 28, 27, 25, 23, 26





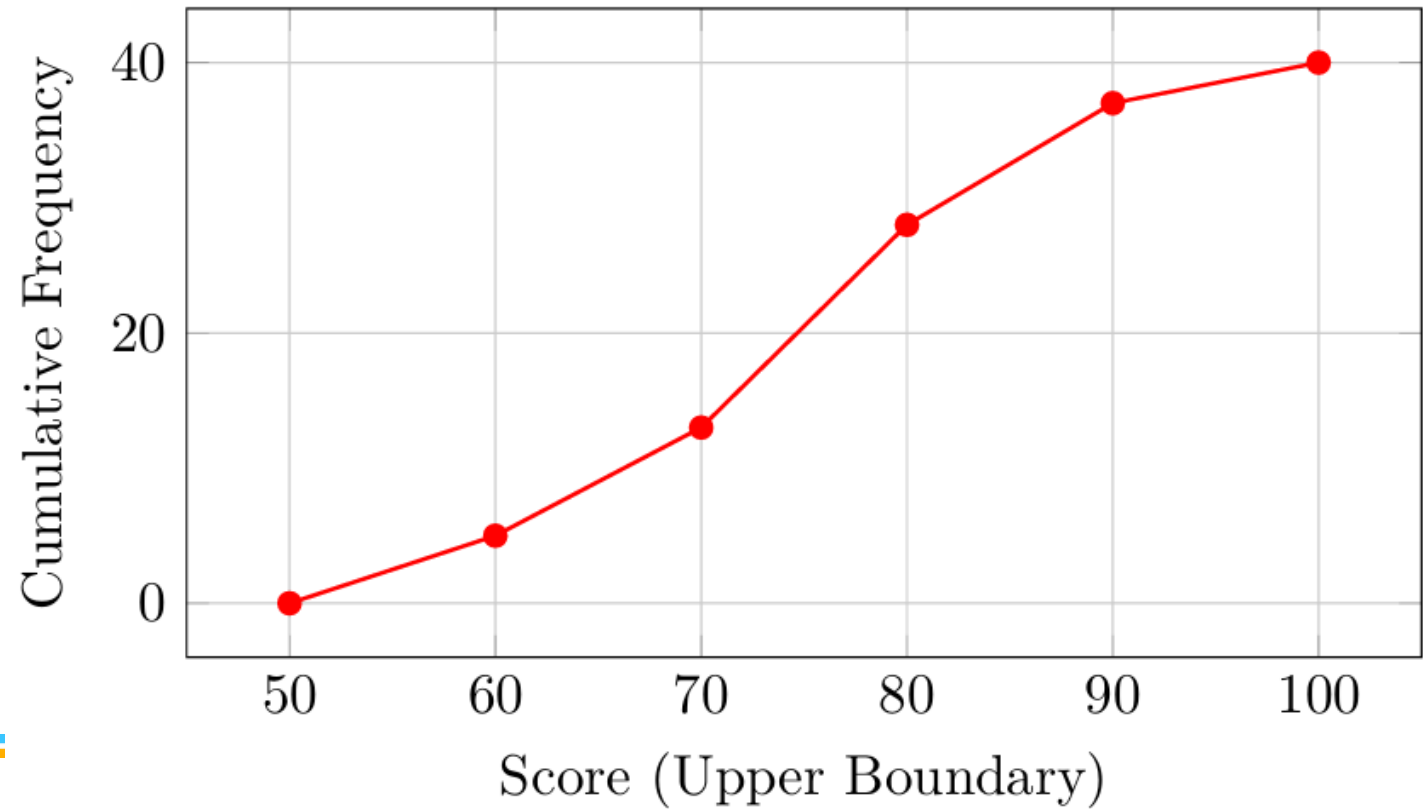
Ogives

An ogive displays cumulative frequencies. For a "less than" ogive, plot upper class boundaries on the x-axis and cumulative frequencies on the y-axis, then connect points with lines.

| Score Range | Frequency | Upper Boundary | Cumulative Freq. |
|-------------|-----------|----------------|------------------|
| 50-60 | 5 | 60 | 5 |
| 60-70 | 8 | 70 | 13 |
| 70-80 | 15 | 80 | 28 |
| 80-90 | 9 | 90 | 37 |
| 90-100 | 3 | 100 | 40 |

Ogives: An ogive displays cumulative frequencies.

| Score Range | Frequency | Upper Boundary | Cumulative Freq. |
|-------------|-----------|----------------|------------------|
| 50-60 | 5 | 60 | 5 |
| 60-70 | 8 | 70 | 13 |
| 70-80 | 15 | 80 | 28 |
| 80-90 | 9 | 90 | 37 |
| 90-100 | 3 | 100 | 40 |





Comparing Quartile Methods



BITS Pilani
DIGITAL
Excellence made yours

np Method: n : total number of observations, p : desired quantiles

Calculate $n \times p$ where p is the quantile (0.25, 0.50, 0.75)

If $n \times p$ is an integer, take the average of the values at that position and the next one.

If $n \times p$ is not an integer, round up to the next integer.

Position Method:

Position =

$$\frac{p(n + 1)}{100}$$

Where p is the percentile (25, 50, 75)



Comparing with Odd n



Test scores for 11 students (sorted): 45, 52, 58, 63, 67, 71, 75, 79, 84, 88, 92

Position Method:

- Q_1 position: $\frac{25(11+1)}{100} = 3 \rightarrow Q_1 = 58$ (3rd value)
- Q_2 position: $\frac{50(11+1)}{100} = 6 \rightarrow Q_2 = 71$ (6th value)
- Q_3 position: $\frac{75(11+1)}{100} = 9 \rightarrow Q_3 = 84$ (9th value)

np Method:

- Q_1 : $11 \times 0.25 = 2.75$ (not integer) \rightarrow round up to 3 $\rightarrow Q_1 = 58$
- Q_2 : $11 \times 0.50 = 5.5$ (not integer) \rightarrow round up to 6 $\rightarrow Q_2 = 71$
- Q_3 : $11 \times 0.75 = 8.25$ (not integer) \rightarrow round up to 9 $\rightarrow Q_3 = 84$

Both Methods give **identical results** for this odd n case!



Comparing with Even n



12 students (Even): 45, 52, 58, 63, 67, 71, 75, 79, 84, 88, 92, 96

Position Method:

- Q_1 position: $\frac{25(12+1)}{100} = 3.25 \rightarrow$ interpolate
 $Q_1 = 58 + 0.25(63 - 58) = 59.25$
- Q_2 position: $\frac{50(12+1)}{100} = 6.5 \rightarrow$ interpolate
 $Q_2 = 71 + 0.5(75 - 71) = 73$
- Q_3 position: $\frac{75(12+1)}{100} = 9.75 \rightarrow$ interpolate
 $Q_3 = 84 + 0.75(88 - 84) = 87$

np Method:

- $Q_1: 12 \times 0.25 = 3$ (integer!)
 $Q_1 = \frac{58+63}{2} = 60.5$
- $Q_2: 12 \times 0.50 = 6$ (integer!)
 $Q_2 = \frac{71+75}{2} = 73$
- $Q_3: 12 \times 0.75 = 9$ (integer!)
 $Q_3 = \frac{84+88}{2} = 86$

Methods give different results



Should method should you use?



BITS Pilani
DIGITAL
Excellence made yours

Data Science Application

Which method should you use?

- The **np method** is what NumPy and most statistical software use by default
- The **position method** is commonly taught in statistics textbooks
- For large datasets ($n > 30$), the differences become negligible
- Always document which method you use for reproducibility

What we have covered

- ✓ Chebyshev's inequality as AI safety net
- ✓ Visualization Techniques
- ✓ Comparing different quartile methods

Thank You

The content provided in this digital training programme has been developed by professors / external vendors. All intellectual property rights in the content, including but not limited to video, text, slides, graphics, and audio materials, are owned exclusively by BITS Pilani ("BITS"), unless otherwise expressly stated. This programme may also include licensed or contributed works from third-party experts, faculty, and collaborating institutions, all of whom are duly acknowledged where appropriate.

All product names, logos, brands, and entity names referenced within this content are the property of their respective trademark owners. The inclusion of such names, logos, or brands does not imply any affiliation with, or endorsement by, these entities.

This material is intended solely for personal learning and academic purposes. Except for content that is explicitly confirmed to be in the public domain, no part of this content may be reproduced, copied, modified, distributed, republished, uploaded, posted, publicly displayed, or transmitted in any form or by any means – electronic, mechanical, photocopying, recording, or otherwise – without the prior express written consent of BITS Pilani or the applicable rights holder.

Any unauthorized use, commercial exploitation, resale, or public sharing of this content in any manner constitutes a violation of applicable intellectual property laws. BITS Pilani reserves the right to pursue all appropriate legal remedies, including but not limited to civil action and criminal prosecution under the Copyright Act, 1957, and other relevant statutes in force in India.

By accessing or using this digital training programme, you agree to comply with the terms set forth in this disclaimer.