Unit 1: Exploring One-Variable Data

Merrick Fanning

September 27, 2025

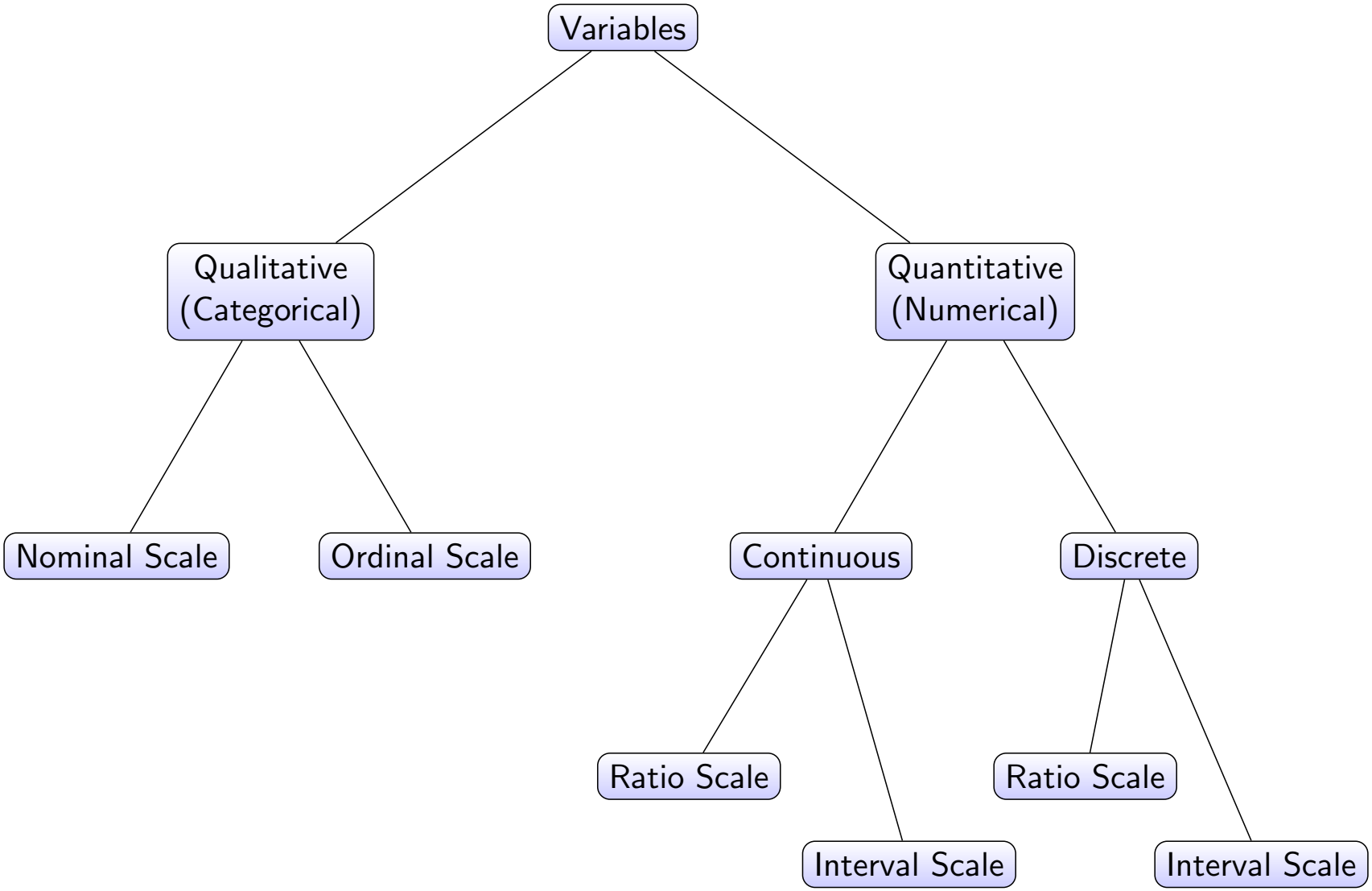
Unit 1 Overview

- Types of Variables
- Representing Categorical Variables with Tables and Graphs
- Representing Quantitative Data with tables and graphs (Dotplots, Histograms, Stemplots, Boxplots)
- **Describing Distributions for Quantitative Variables (SOCS)**
- Measures of Center and Spread
- Graphical Representations of Summary Statistics
- Comparing Distributions.
- Introduction to the Normal Distribution
- Effects of Linear Transformations on Data
- Problems Involving Mean/Median/Mode.

Questions

1. What is a random variable?
2. How do we classify random variables?
3. What is the context for random variables in statistics?

Types of Variables



Interval vs Ratio Scales

Feature	Interval Scale	Ratio Scale
Type of variable	Quantitative	Quantitative
Ordered values?	Yes	Yes
Meaningful differences?	Yes	Yes
Has a true zero?	**No** (arbitrary zero)	**Yes**
Ratios are meaningful?	**No**	**Yes**
Can say "twice as much"?	No	Yes
Example units	Celsius, IQ, calendar years	Kelvin, weight, height, age

Table: Comparison of Interval and Ratio Scales

Note: It's very unlikely you will be asked questions about scales anywhere in AP Statistics.

Dataset #1: Penguins

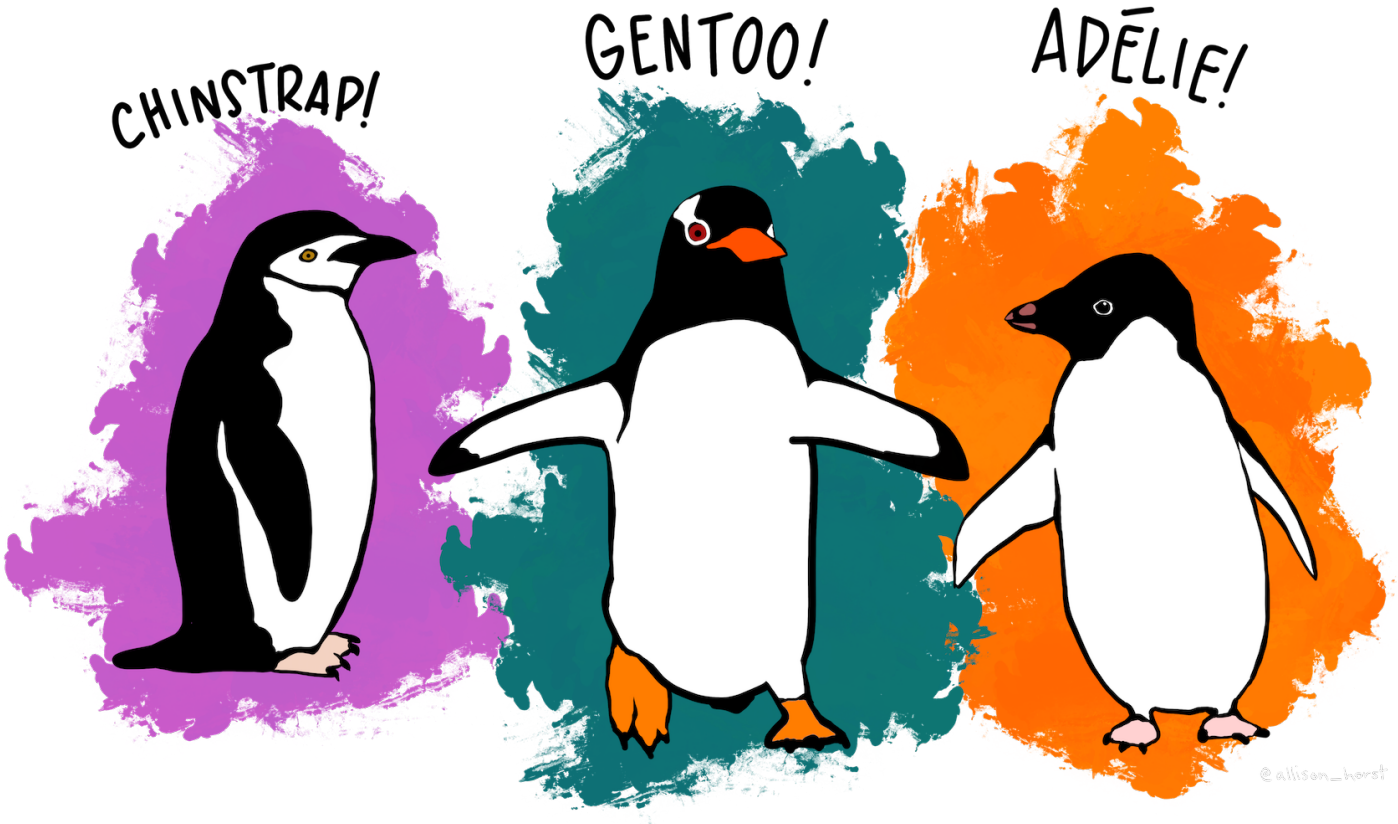


Figure: Palmer Archipelago Penguin Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Sample Penguin Data

Table: Palmer Archipelago Penguin Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Species	Island	Bill Len.	Bill Dep.	Flipper Len.	Mass	Sex
Adelie	Torgersen	39.1	18.7	181	3750	Male
Adelie	Torgersen	39.5	17.4	186	3800	Female
Chinstrap	Dream	46.5	17.9	192	3500	Male
Gentoo	Biscoe	50.0	15.3	220	5550	Female
Adelie	Dream	37.8	18.3	174	3400	Female
Chinstrap	Dream	45.2	17.2	193	3650	Male
Gentoo	Biscoe	48.7	14.1	210	5050	Male
Adelie	Torgersen	38.2	18.1	180	3700	Female
Chinstrap	Dream	47.1	16.8	195	3600	Female
Gentoo	Biscoe	49.5	15.0	217	5400	Male
Adelie	Dream	40.3	18.5	185	3900	Male
Chinstrap	Dream	44.8	17.6	190	3550	Female

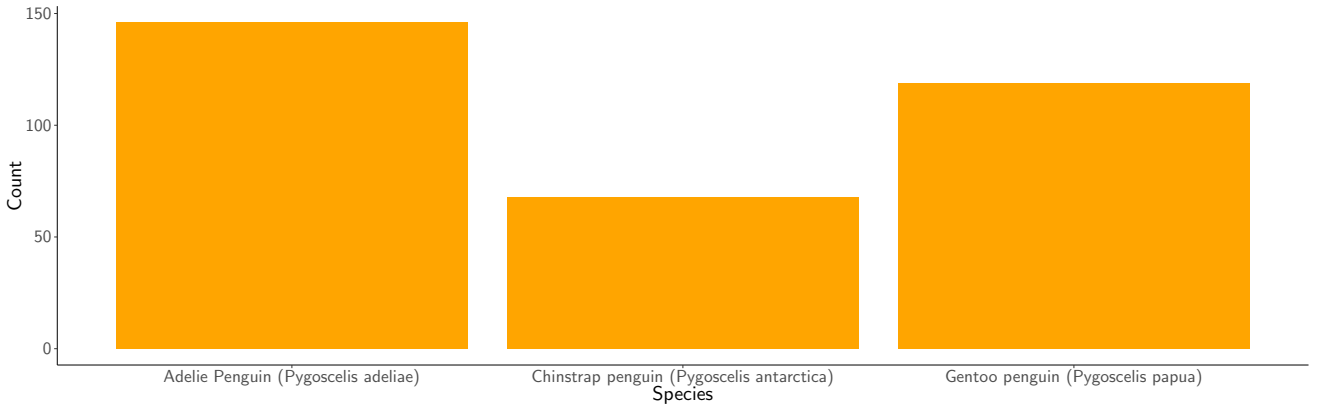
One Categorical Variable: Tables and Graphs

1. How would you represent the 'Species' random variable with a table?

Species	**Count**
Adelie Penguin (*Pygoscelis adeliae*)	146
Chinstrap Penguin (*Pygoscelis antarctica*)	68
Gentoo Penguin (*Pygoscelis papua*)	119

Table: Distribution of Penguin Species in the Dataset

2. What is *frequency*? What about *relative frequency*?
3. How would you represent the 'Species' random variable with a graph?



Summarizing One Variable Quantitative Data

Consider the dataset listed below:

$$\{2, 4, 6, 8\}$$

- How do we measure the center of the data?
  1. Median
  2. Mean
  3. Mode
- How do we measure the spread of the data?
  1. Range
  2. Variance
  3. Standard Deviation
  4. Quartiles (Deciles? Percentiles?)
  5. Interquartile Range (IQR)
  6. Mean Absolute Deviation
- Which of these observations might be *sensitive* to outliers?
  1. Mean
  2. Variance

Outliers

Consider the dataset describing some test scores below:

$$\{41, 41, 81, 82, 85, 86, 86, 86, 89, 88, 100\}$$

1. Which observations seem to be 'extreme'
2. What is the rule for calculating outliers?
   1. $x < Q1 - 1.5(IQR)$ or $x > Q3 + 1.5(IQR)$
   2. 2 or more standard deviations away from the mean.
3. How do you calculate outliers using a calculator?

Symbols and Formulas to Know

Measures of Center:

- $\overline{x}$: Sample mean $\quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- μ: Population mean (typically unknown)
- Median: Middle value (50th percentile)
- Mode: Most frequently occurring value(s)

Measures of Spread:

- s^2: Sample variance $\quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- σ^2: Population variance $\quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- s: Sample standard deviation $\quad s = \sqrt{s^2}$
- σ: Population standard deviation
- Range: $\max - \min$
- IQR: Interquartile Range $= Q_3 - Q_1$

One Numeric Variable: Tables and Graphs

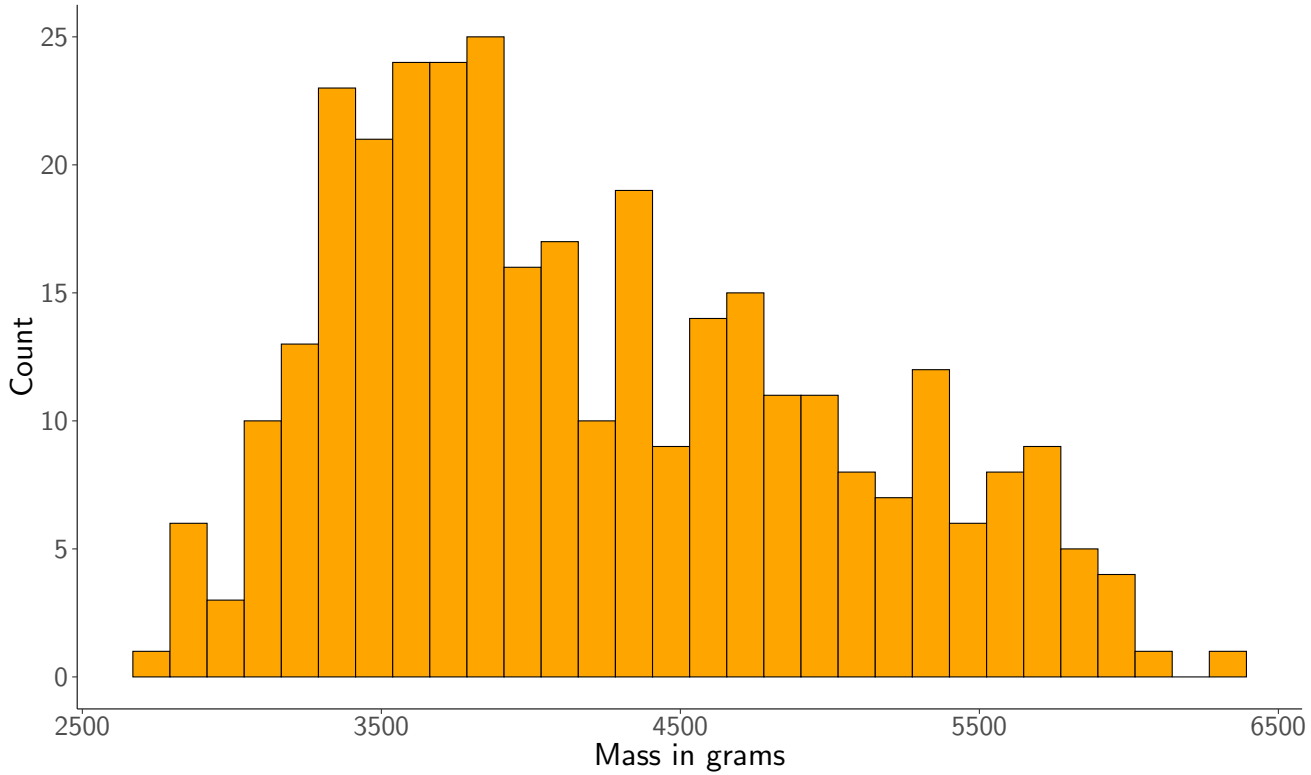
1. How could you represent the 'Mass' variable for the penguins data with a table?

Body Mass Range (g)	Count
$(2700, 3420]$	56
$(3420, 4140]$	121
$(4140, 4860]$	80
$(4860, 5580]$	54
$(5580, 6300]$	22

Table: Penguin Body Mass Binned into 5 Equal Intervals

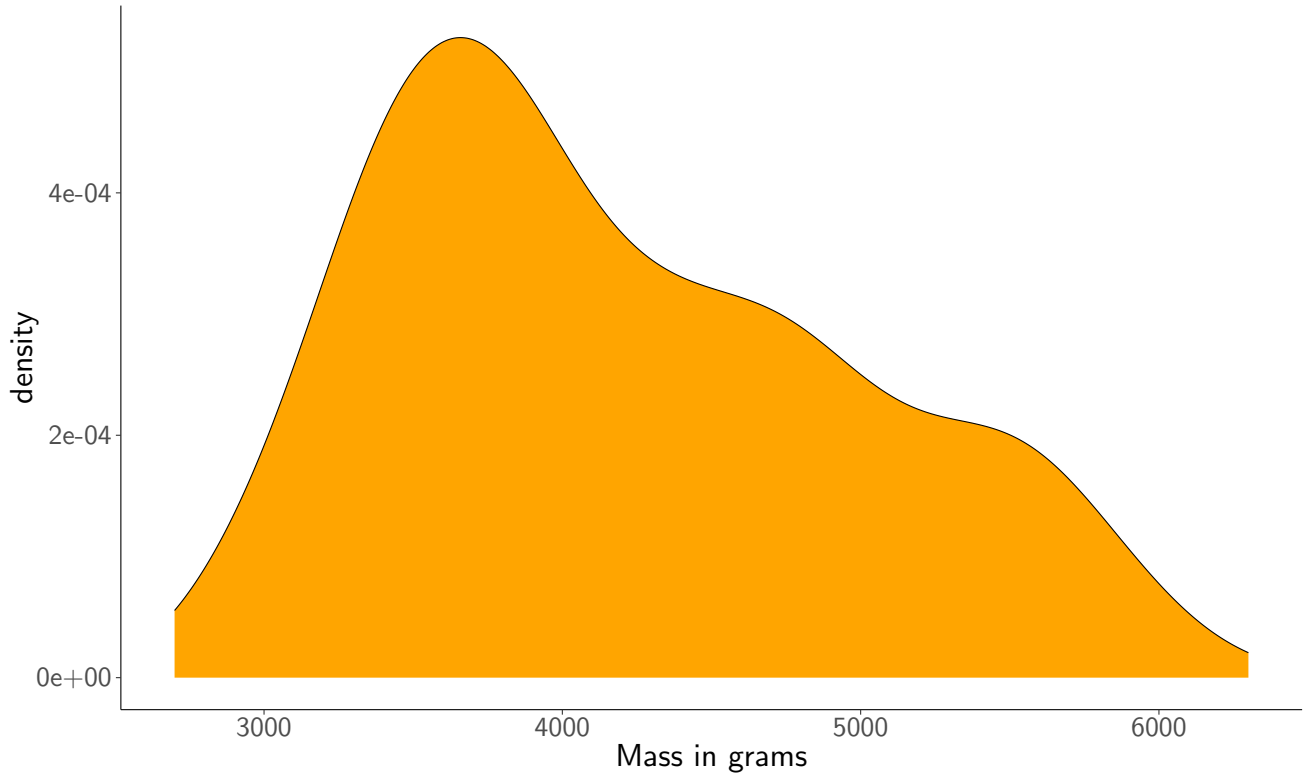
2. How could you represent the 'Mass' variable for the penguins data with a graph?

Histograms

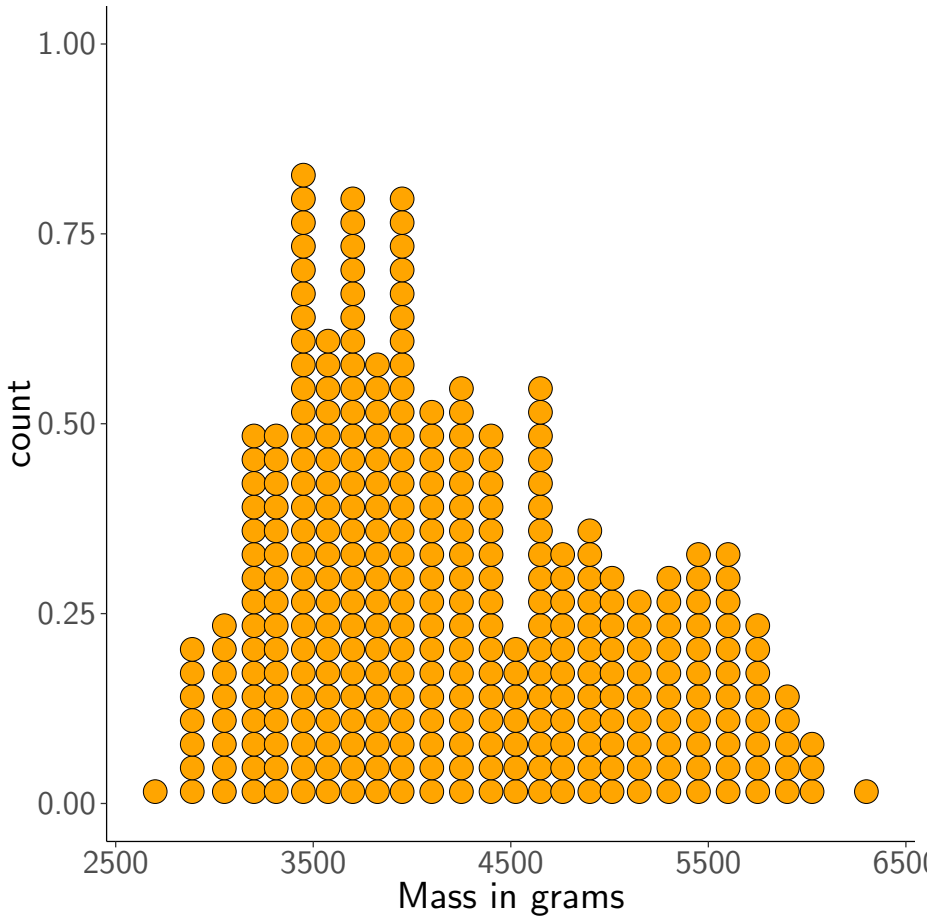


- What are some advantages to histograms?
- What is the difference between a histogram and a barchart?
- What does changing the size of bins do?

Density Plots

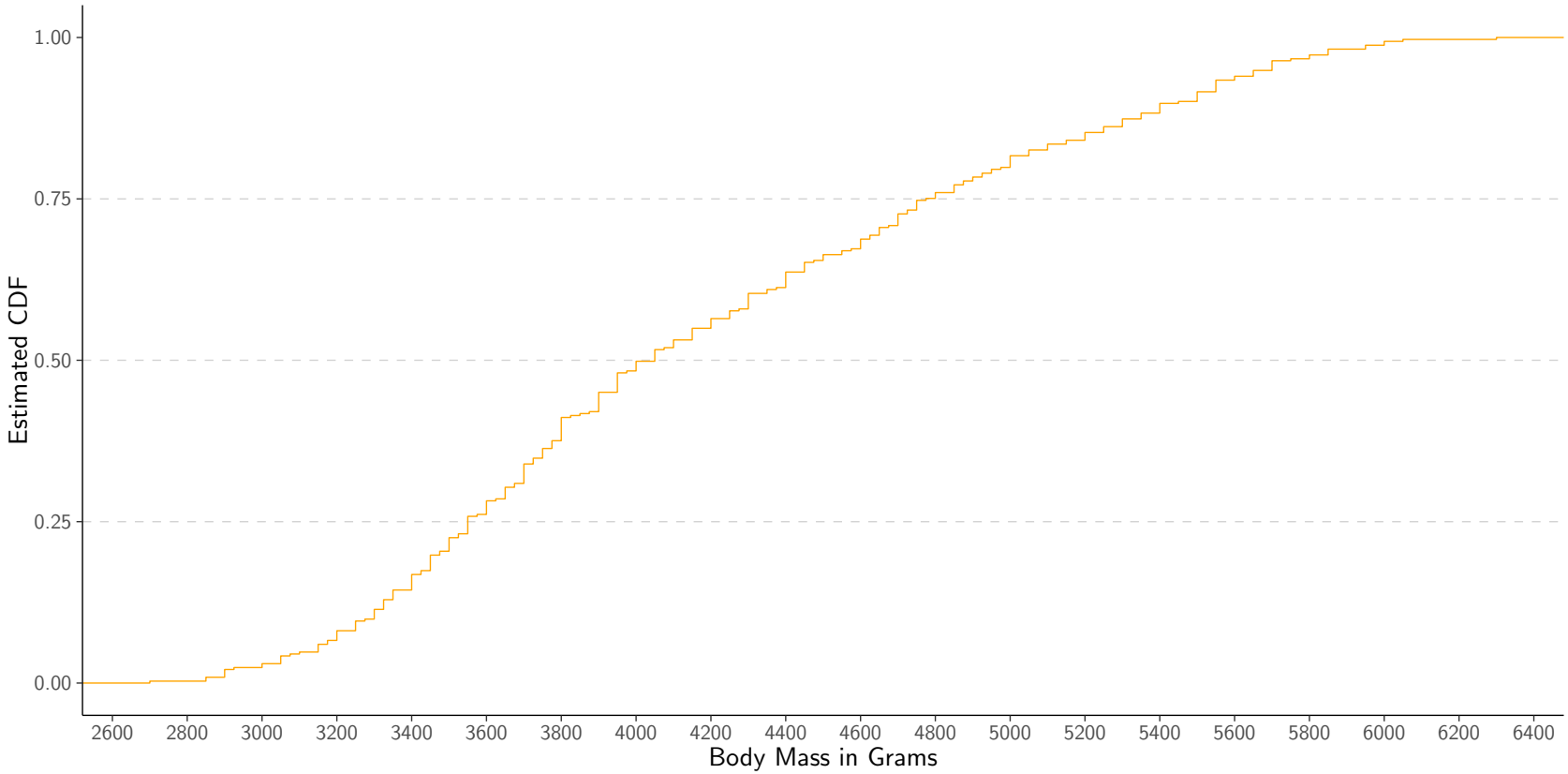


Dot Plots



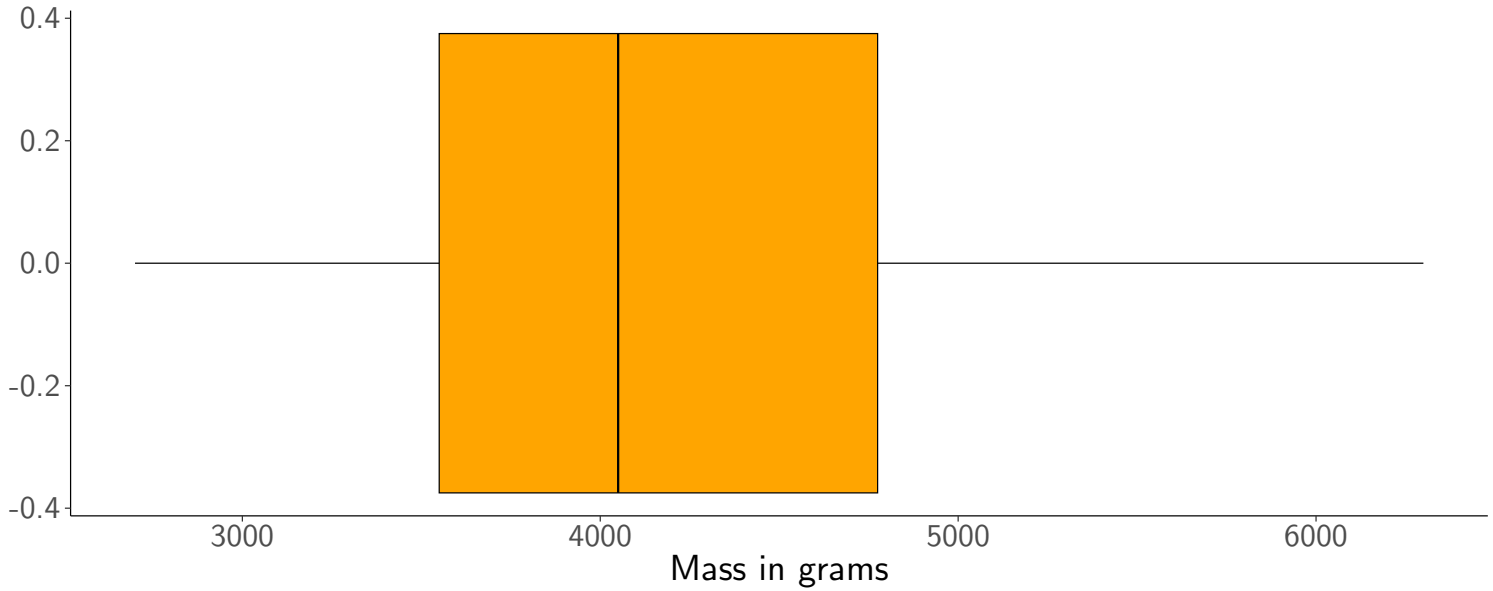
- Why would a dot plot be better or worse than a histogram?

Cumulative Distribution Plots



1. What is the median of the distribution? Various quartiles?
2. What is the relationship between CDF plots and Histograms/Density plots?
3. Draw the CDF for skewed right, normal, and skewed left data.

Boxplots (Box and Whiskers Plots)



- Why would a historgram be used rather than a box plot?
- How are outliers added to a box plot?

Stem-and-Leaf Plots

Distribution for test scores in points

Stem	Leaf
5	2 5 7
6	1 3 4 8 9
7	0 1 3 5 5 8
8	2 4 6
9	1 3

Key: 7 | 5 means 75 points

Note: Don't forget to add units, label, and key!

Describing the Distribution of a Numeric Variable: **SOCS**

When describing the distribution of a numeric random variable, use the SOCS framework:

- **S** **Shape** - What is the overall form of the distribution?
  - Symmetric, skewed left, skewed right
  - Unimodal, bimodal, uniform
- **O** **Outliers and Unusual Features** - Are there any values or patterns that stand out?
  - Obvious gaps, clusters, or isolated points
  - Extreme values (high or low)
  - Anything that doesn't follow the overall pattern
- **C** **Center** - Where is the middle of the data?
  - Median (if skewed or with outliers). Why not use mean?
- **S** **Spread** - How spread out is the data?
  - Range. Why not use standard deviation?

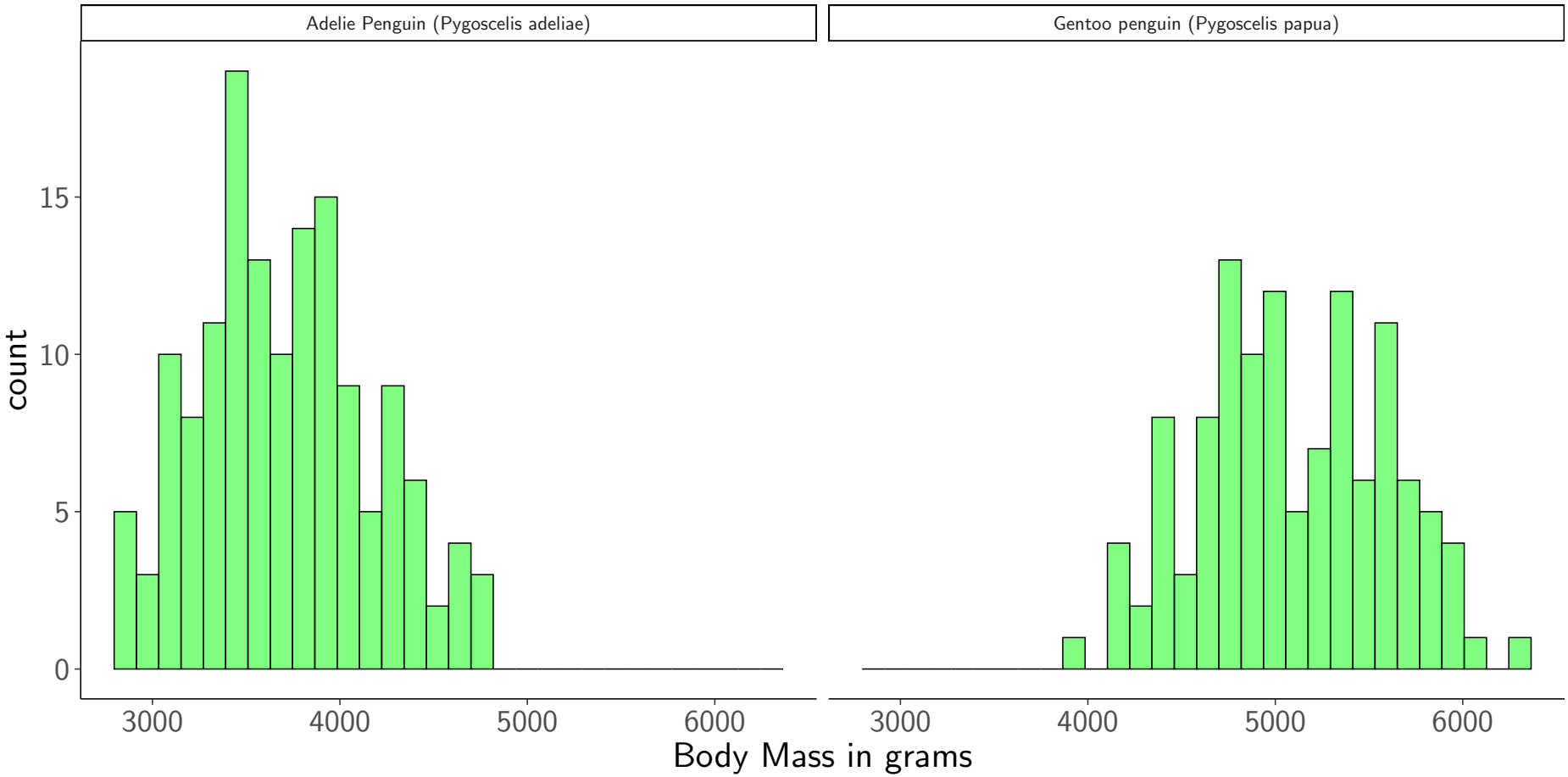
Always describe SOCS in context and refer to visual displays (like histograms or

Describing a Distribution (SOCS Example)

SOCS Description:

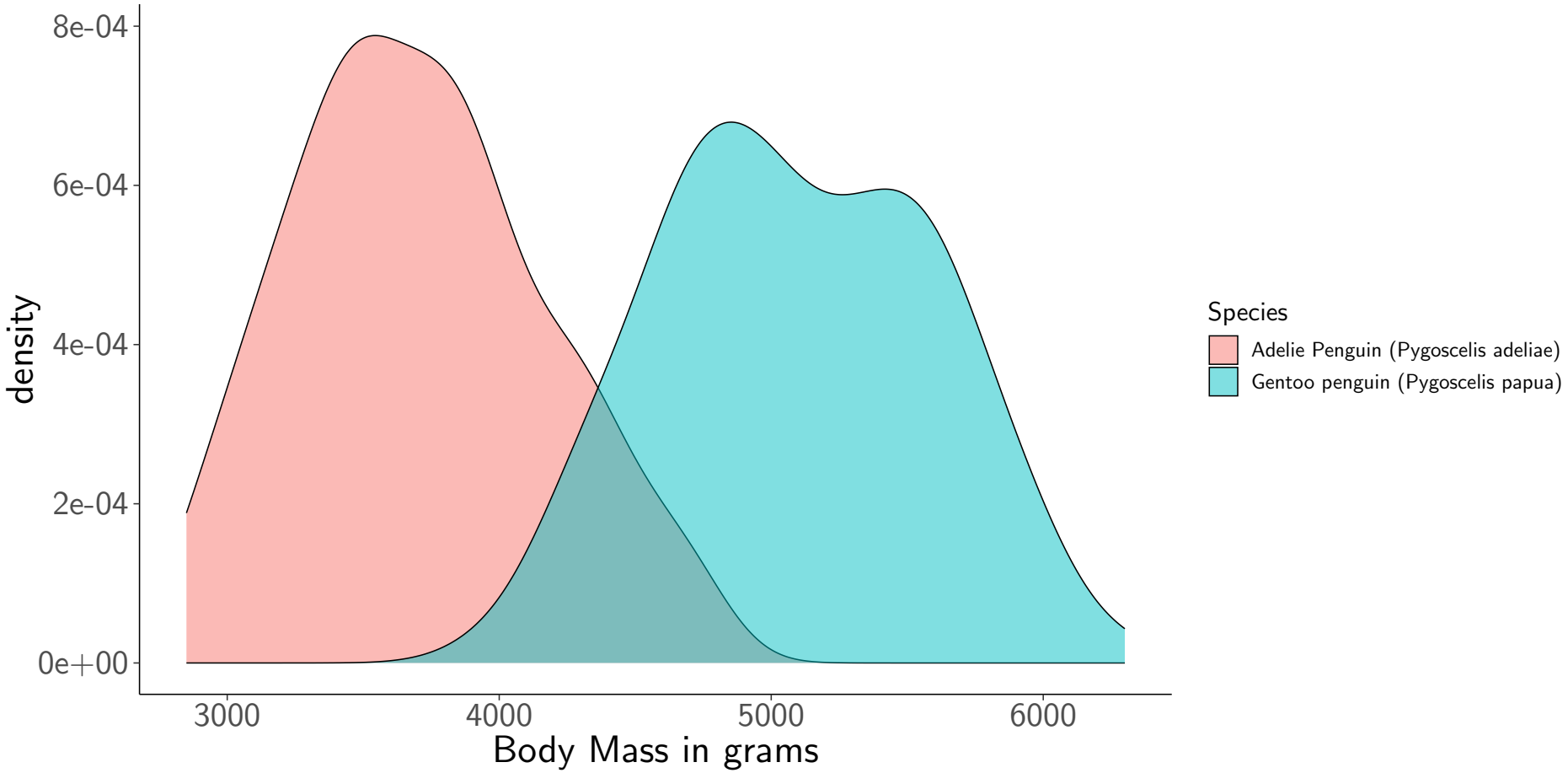
- **Shape:** The distribution of sampled lego prices in Dollars is skewed to the right.
- **Outliers and Unusual Features:** There are several outliers with prices higher than $1350. there is a large gap from $170 to $280.
- **Center:** The median appears to be around $30.
- **Spread:** The range is $230.

Comparing Distributions (SOCS Example)

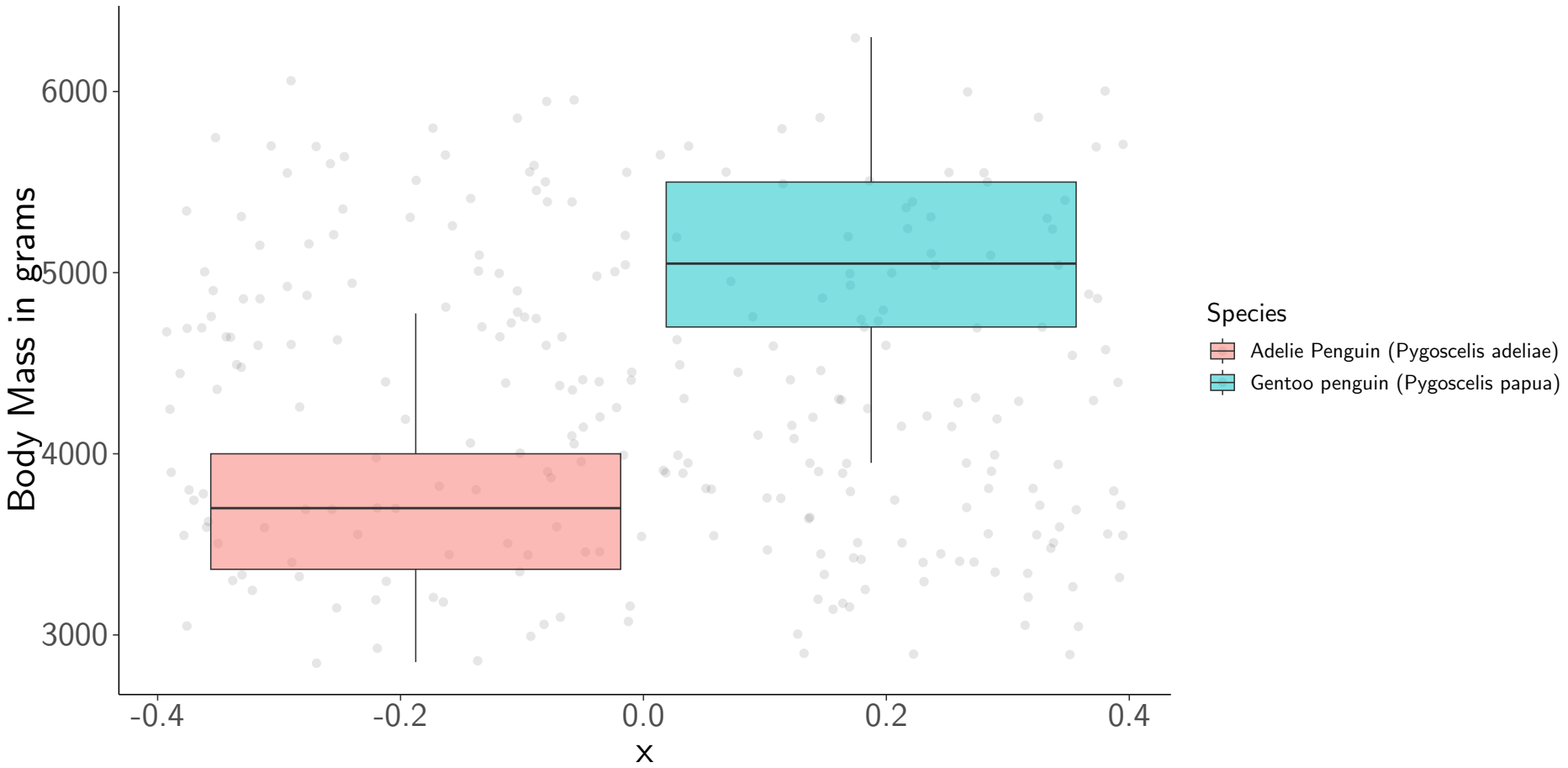


- To compare distributions use SOCS Description Frame work
- Don't forget to **compare**

Comparing Distributions



Comparing Distributions



Chebyshev's Theorem

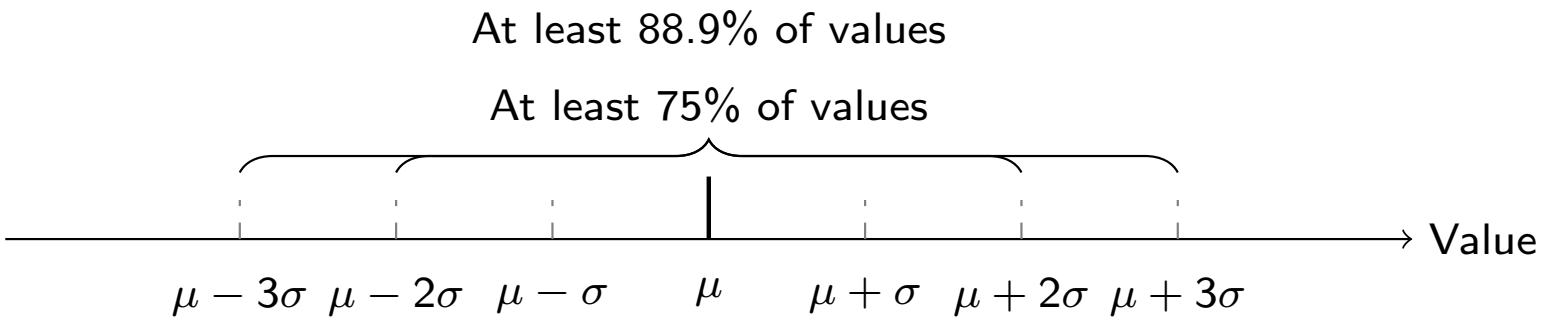
Mathematical Definition:
For any real number $k > 1$, and any distribution with finite mean μ and standard deviation σ:

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

Interpretation:
No matter the shape of the distribution, at least $1 - \frac{1}{k^2}$ of the data values must lie within k standard deviations of the mean.

Examples:

- $k = 2$: At least 75% of values lie within $\mu \pm 2\sigma$
- $k = 3$: At least 88.9% of values lie within $\mu \pm 3\sigma$



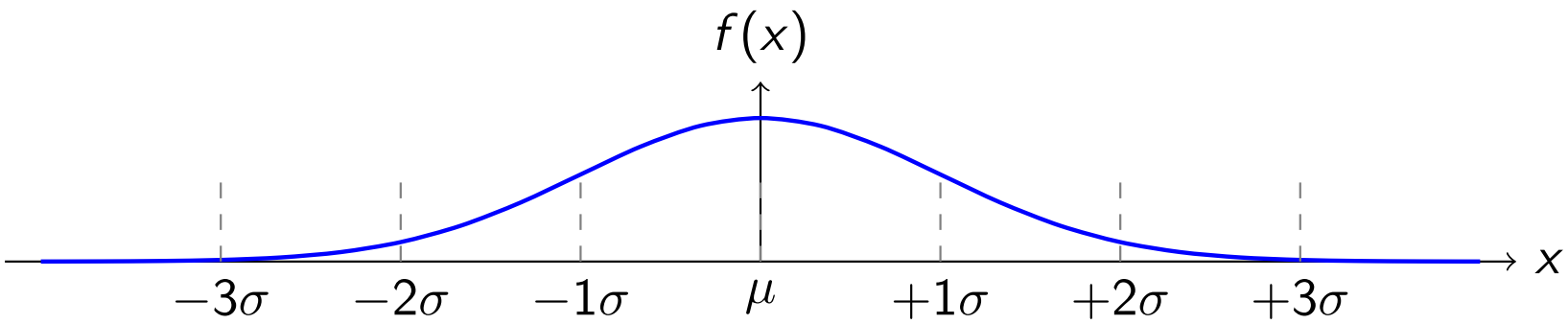
The Normal Distribution

Definition: The normal distribution is a symmetric, bell-shaped distribution that describes many natural phenomena. It is defined by its mean μ and standard deviation σ.

Probability Density Function (PDF):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Standard Normal Curve:



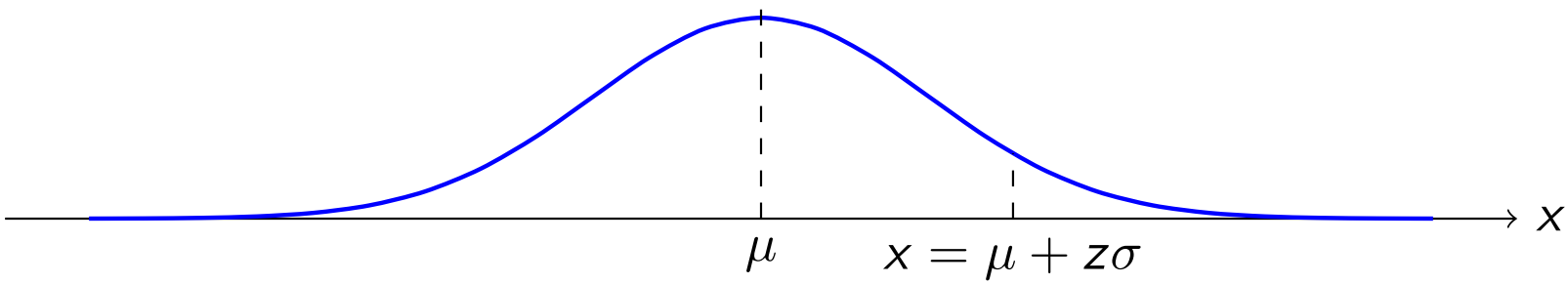
- What happens when you vary μ?
- What happens when you vary σ?

Z-Scores and Standardization

Definition: A **z-score** tells you how many standard deviations a data value is from the mean.

$$z = \frac{x - \mu}{\sigma} \quad \Rightarrow \quad x = \mu + z\sigma$$

- If $z > 0$: the value is above the mean
- If $z < 0$: the value is below the mean
- If $|z| > 2$: the value may be considered unusually far from the mean



1. Why might we want to standardize?
2. Can you think of an example where a z-score would be more useful than the actual value?

Solving for Standard Deviation Using Percentiles

In a university biology course, students' exam scores are approximately normally distributed.

- A student who scored **92** was in the **90th percentile**
- Another student who scored **76** was in the **42nd percentile**

Determine the standard deviation σ of the exam scores.

Step 1: Convert percentiles to z-scores

$$\text{90th percentile} \Rightarrow z_1 \approx 1.28$$
$$\text{42nd percentile} \Rightarrow z_2 \approx -0.20$$

Step 2: Set up the system

$$92 = \mu + 1.28\sigma \quad (1)$$
$$76 = \mu - 0.20\sigma \quad (2)$$

Step 3: Subtract the equations

$$(92 - 76) = (1.28 + 0.20)\sigma \Rightarrow 16 = 1.48\sigma \Rightarrow \sigma = \boxed{10.81}$$

The Empirical Rule (68-95-99.7 Rule)

The Empirical Rule:
For a distribution that is approximately normal:

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

Question How might you estimate the standard deviation of roughly bell shaped data using the range?

Normal Probability Plots

A **normal probability plot** helps assess whether a data set is approximately normally distributed.

How It's Made:

1. Order the data from smallest to largest.
2. Determine each value's percentile. Often done using Blom's formula.
3. Calculate the theoretical z-score (expected under a normal distribution) for each percentile.
4. Plot the actual data values against these theoretical $z-$scores.

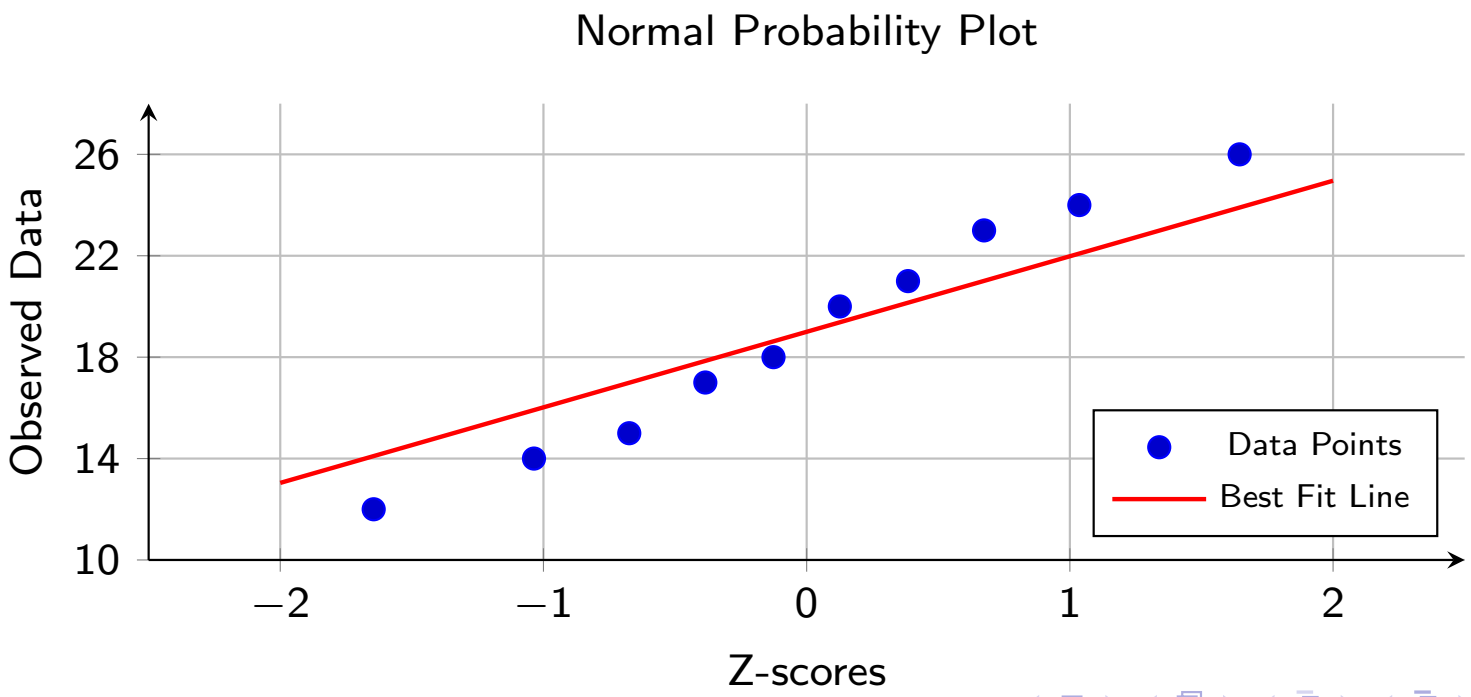
How to Interpret:

- If the plot is roughly a straight line, the data are approximately normal.
- Systematic curves or bends suggest skewness or non-normality:
  - S-shape: heavy tails or skewness
  - Curve up/down at ends: non-normal tails

Summary:

Normal Probability Plot Example

Index i	1	2	3	4	5	6	7	8	9	10
Data d_i	12	14	15	17	18	20	21	23	24	26
Percentile $p_i = \frac{i-0.5}{n}$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Z-score	-1.645	-1.036	-0.674	-0.385	-0.126	0.126	0.385	0.674	1.036	1.645



Blom's Method for Percentiles in Normal Probability Plots

Goal: Assign a percentile to each data point in a way that spreads them appropriately across the standard normal distribution.

Blom's Formula:

$$p_i = \frac{i - 0.5}{n}$$

Where:

- i is the rank of the data point ($1 =$ smallest, $n =$ largest)
- n is the total number of data points
- p_i is the assigned percentile (used to look up a z-score)

Why this works:

- It avoids assigning 0% to the first data point or 100% to the last - which would imply infinite z-scores.
- It spreads points symmetrically around the center (mean) of the normal distribution.
- It better matches the expected order statistics under a normal distribution.

Effect of Adding a Constant to All Data Values

Let a dataset of sample data be defined as:

$$D = \{d_1, d_2, \ldots, d_n\}$$

Let $c \in \mathbb{R}$ be a constant. When adding a constant to all values the transformed dataset is:

$$D' = \{d_1 + c, d_2 + c, \ldots, d_n + c\}$$

1. What effect will this have on the sample mean?
2. What effect will this have on the sample variance?
3. What effect will this have on the sample standard deviation?

Effect of Multiplying All Data Values by a Constant

Let a dataset of sample data be defined as:

$$D = \{d_1, d_2, \ldots, d_n\}$$

Let $a \in \mathbb{R}$ be a constant. When multiplying all values by a constant, the transformed dataset is:

$$D' = \{a \cdot d_1, a \cdot d_2, \ldots, a \cdot d_n\}$$

1. What effect will this have on the sample mean?
2. What effect will this have on the sample variance?
3. What effect will this have on the sample standard deviation?

Effect of Linear Transformations on All Data Values

Let a dataset of sample data be defined as:

$$D = \{d_1, d_2, \ldots, d_n\}$$

Let $a, b \in \mathbb{R}$. When each data value is transformed by multiplying by a and adding b, the new dataset is:

$$D' = \{a \cdot d_1 + b,\ a \cdot d_2 + b, \ldots, a \cdot d_n + b\}$$

1. What effect will this have on the sample mean?
2. What effect will this have on the sample variance?
3. What effect will this have on the sample standard deviation?

Effect of a Non-Uniform Transformation on Mean and Variance

Let a dataset from a sample be defined as:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Now define a transformation where each data value is increased by its index:

$$x_i' = x_i + i \quad \text{for all } i = 1, 2, \ldots, n$$

Questions:

1. What effect does this transformation have on the sample mean?
2. What effect does this transformation have on the sample variance?

Solution: Effect of Adding i to Each Data Value

Recall the transformation:

$$d_i' = d_i + i \quad \text{for } i = 1, 2, \ldots, n$$

Effect on the Mean:

$$\bar{d}' = \frac{1}{n}\sum_{i=1}^{n} d_i' = \frac{1}{n}\sum_{i=1}^{n}(d_i + i) = \bar{d} + \frac{1}{n}\sum_{i=1}^{n} i = \bar{d} + \frac{n+1}{2}$$

So the mean increases by $\frac{n+1}{2}$

Effect on the Variance:

$$\text{Var}(d_i') = \text{Var}(d_i + i)$$

Assuming the d_i's are independent of their indices i:

$$\text{Var}(d_i') = \text{Var}(d_i) + \text{Var}(i)$$

$$\text{Var}(i) = \text{Var}(1, 2, \ldots, n) = \frac{(n^2 - 1)}{12}$$