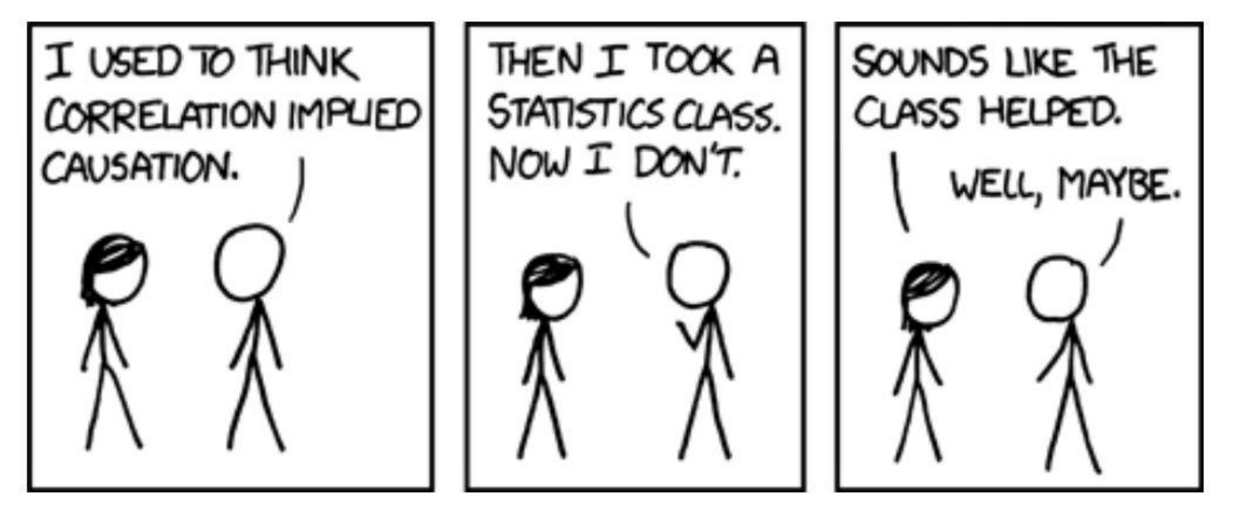
STATISTICS AND DATA

Assistant Professor Gwendolyn Eadie

David A. Dunlap Dept. of Astronomy & Astrophysics / Dept. of Statistical Sciences

University of Toronto, Toronto, Ontario, Canada

May 5, 2025


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Statistics

Study of uncertainty, estimating and summarizing properties about **data**, learning from **data**, analysing **data**, collecting **data** …

It can be used to infer facts, make predictions, and make recommendations, **based on data**.

Nearly all fields that collect data use statistics: astronomy, biology, chemisty, environmental science, forestry, geophysics, law, politics, sports analytics, etc.

What's the recurring theme?

Data!

Broad categories of data

Quantitative

- Numerical
- e.g., height, age, time since an event, blood pressure, brightness of a star, etc.

Categorical

- Can be grouped into a category, type, or quality
- e.g., letter grade, month of birthday, type of galaxy, type of treatment, etc.

Ordinal

- Have a natural order
- Differences between two values may not be meaningful

How we visualize and summarize data depends on the type of data we have

Types of data

Spatial data

Time Series Data

Spatio-temporal data

Count data

Multivariate data

Combinations of these!

How we visualize and summarize data depends on the type of data we have

Terminology: a "population" versus a "sample"

Population

- The true, underlying distribution for some quantity
- E.g., the distribution of heights of people all over the world

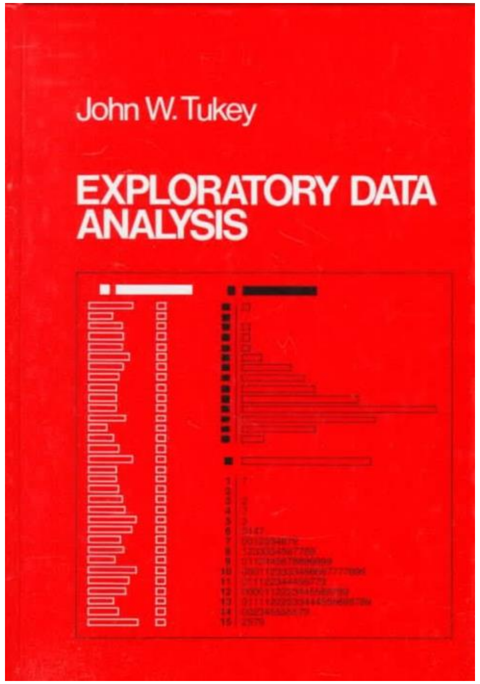
Sample

- A sample drawn from some distribution
- E.g., randomly select 100 people from around the world and measure their heights
- Will not be exactly like the population because of randomness

Statistics can be used to try to understand the underlying population, when all you have is a sample

“Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.”

Tukey, J.W., “The Future of Data Analysis”, *The Annals of Mathematical Statistics*, Mar., 1962, Vol. 33, No. 1 (Mar., 1962),
pp. 1-67



Exploratory Data Analysis (EDA)

What is Exploratory Data Analysis and what is it for?

- explores data to better understand their characteristics
- uses visual methods and summary statistics
- help formulate possible hypotheses, generate questions
- help us understand/identify outliers, obvious errors, and quirks in the data
- “check” to see if a particular data analysis technique is appropriate for the particular data set
- reveal additional information not directly related to the research question

Exploratory vs. Confirmatory

Exploratory

- Explore the data
- Generate hypotheses or questions
- Refine scientific questions
- Can be an iterative cycle

Confirmatory

- Hypothesis/model generation
- Data collection and experiment
- Hypothesis testing, parameter inference, etc.

Wickham & Grolemund, *R for Data Science,* O'Reilly,

There is no rule about which questions you should ask to guide your research. However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1. What type of variation occurs within my variables?
2. What type of covariation occurs between my variables?

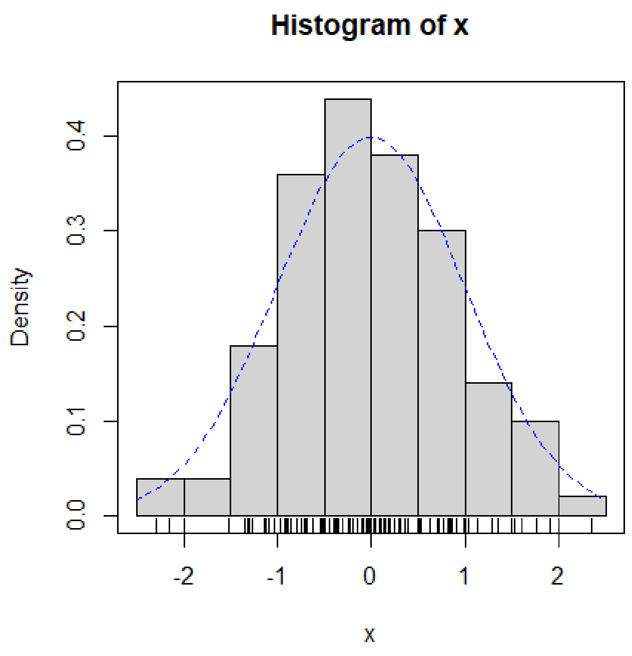
Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

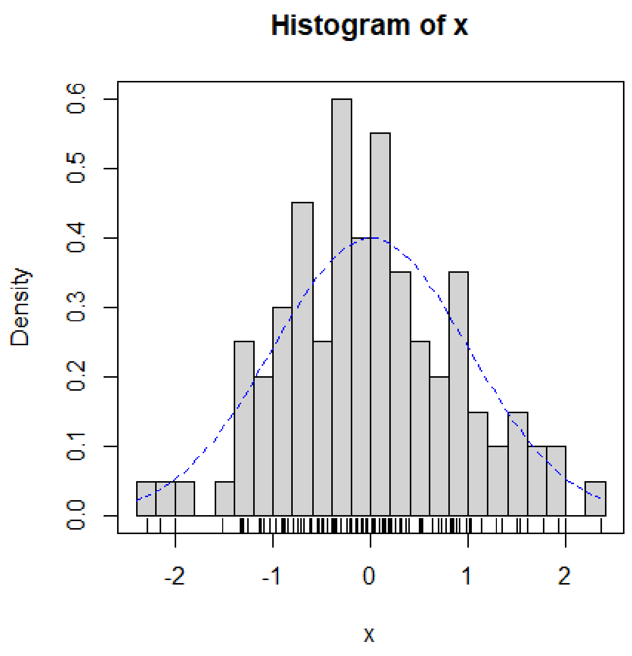
What are some EDA techniques/visualizations of data?

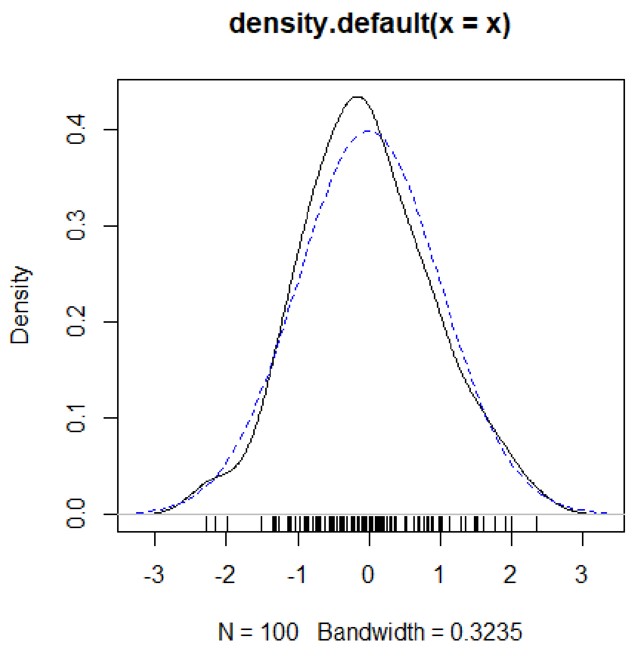
- Histograms (or density plots)
- Scatter plots
- Conditioning plots
- Pairs plot (pair-wise scatterplot)
- Tukey’s 5-number summary
- Boxplots, violin plots
- Mosaic Plots

Histograms & Kernel Density Estimators (KDE)

- *Estimators* of the underlying density distribution
- Histogram pros & cons
  - Dependent on bin sizes, breakpoints
- KDE pros & cons
  - does not bin the data
  - *Bias-variance* trade-off when choosing the bandwidth parameter



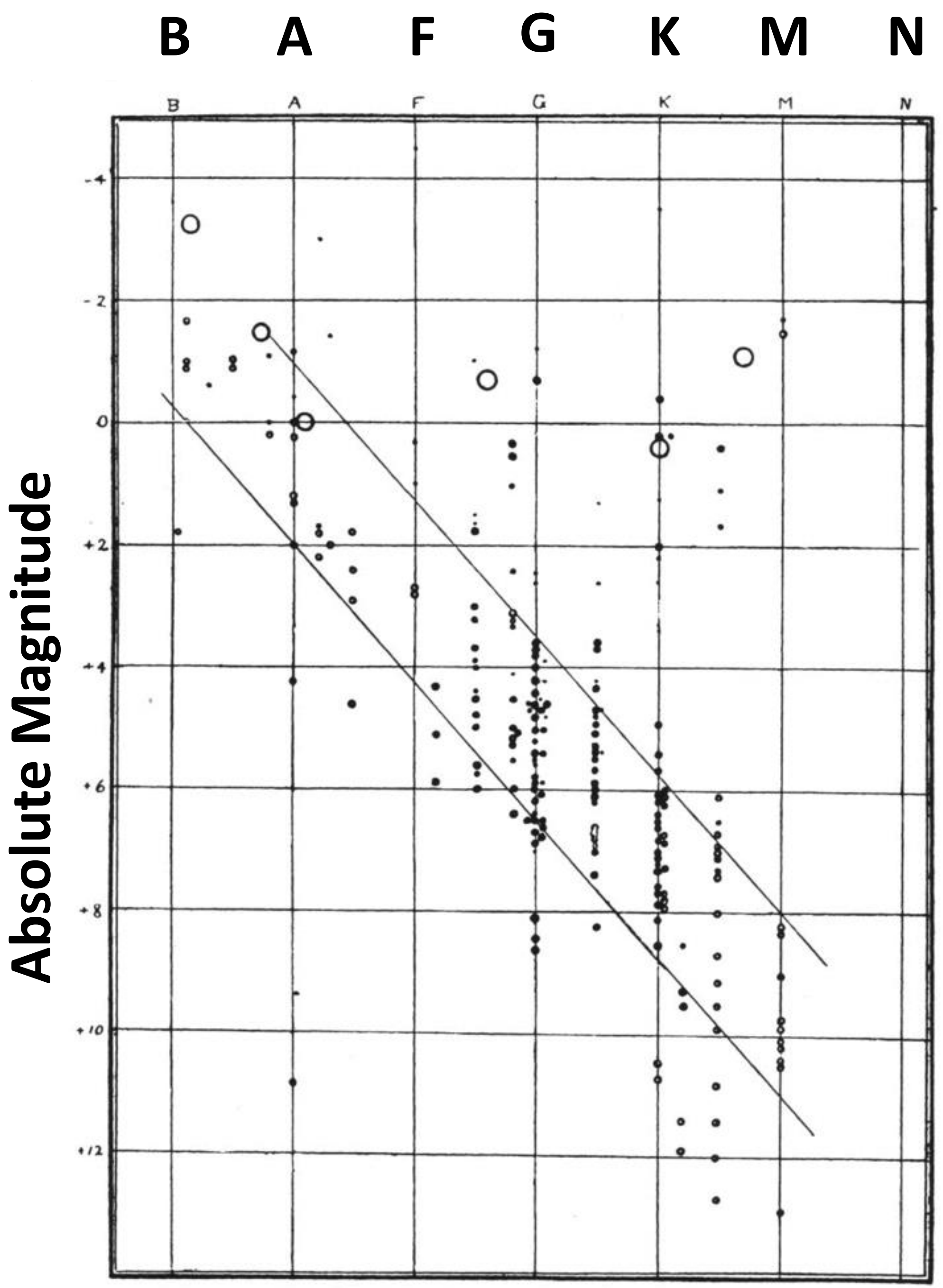




Can you think of EDA techniques that have been transformative in astronomy?

Hertzsprung-Russell Diagram

- Observational H-R diagram
- Theoretical H-R diagram



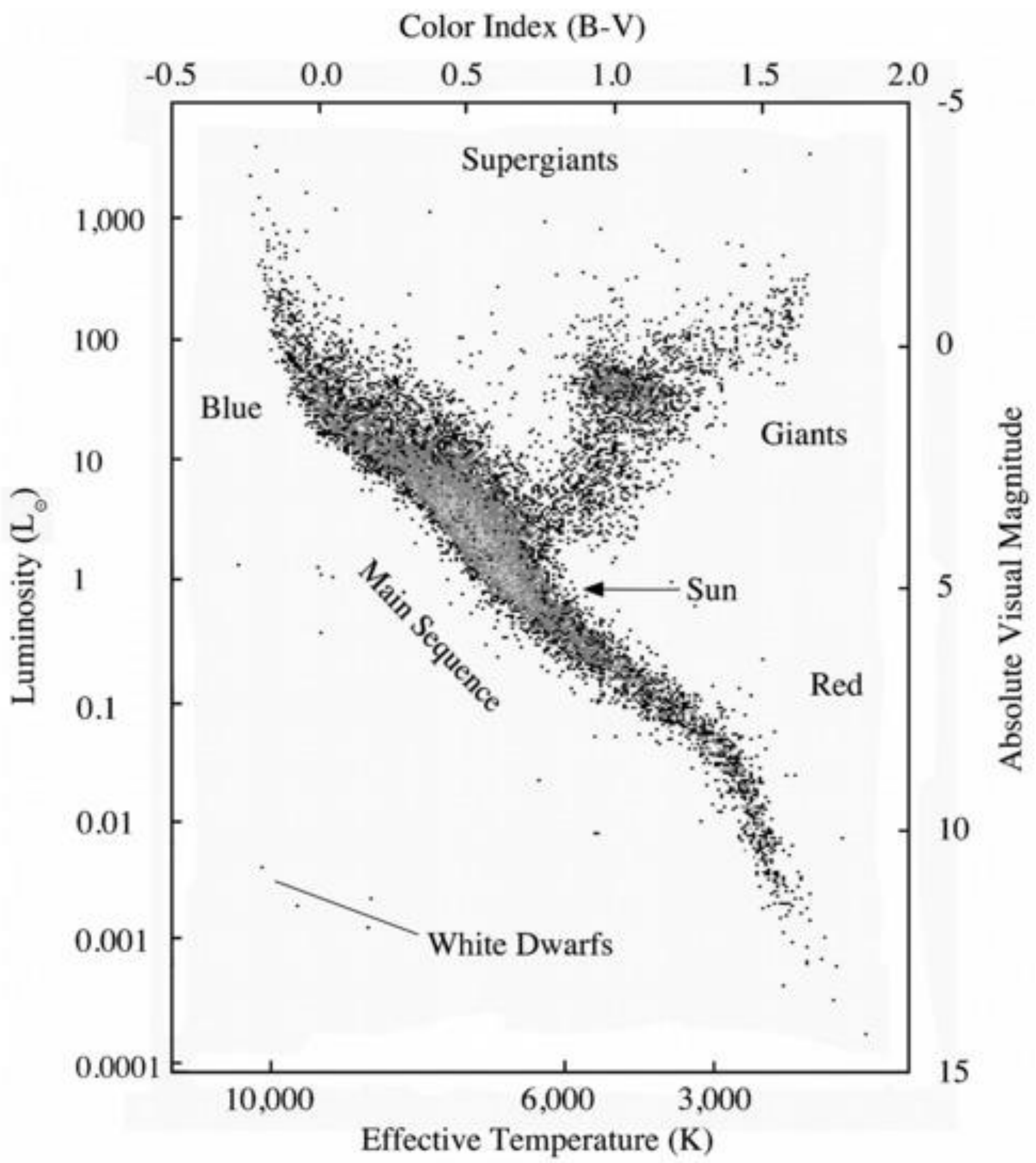
Hertzsprung and Russell independently discovered a relationship between the spectral type of stars and their absolute brightness.

← One of the first *observational* H-R diagrams, appearing in Russell(1914) *Nature*, 93, 252.

→ What type of exploratory data analysis made this possible?

- Looking at the data!
- Scatter plot

Data from **Hipparcos** Satellite



https://ase.tufts.edu/cosmos/print_images.asp?id=49

Image Credit:
Michael Perryman

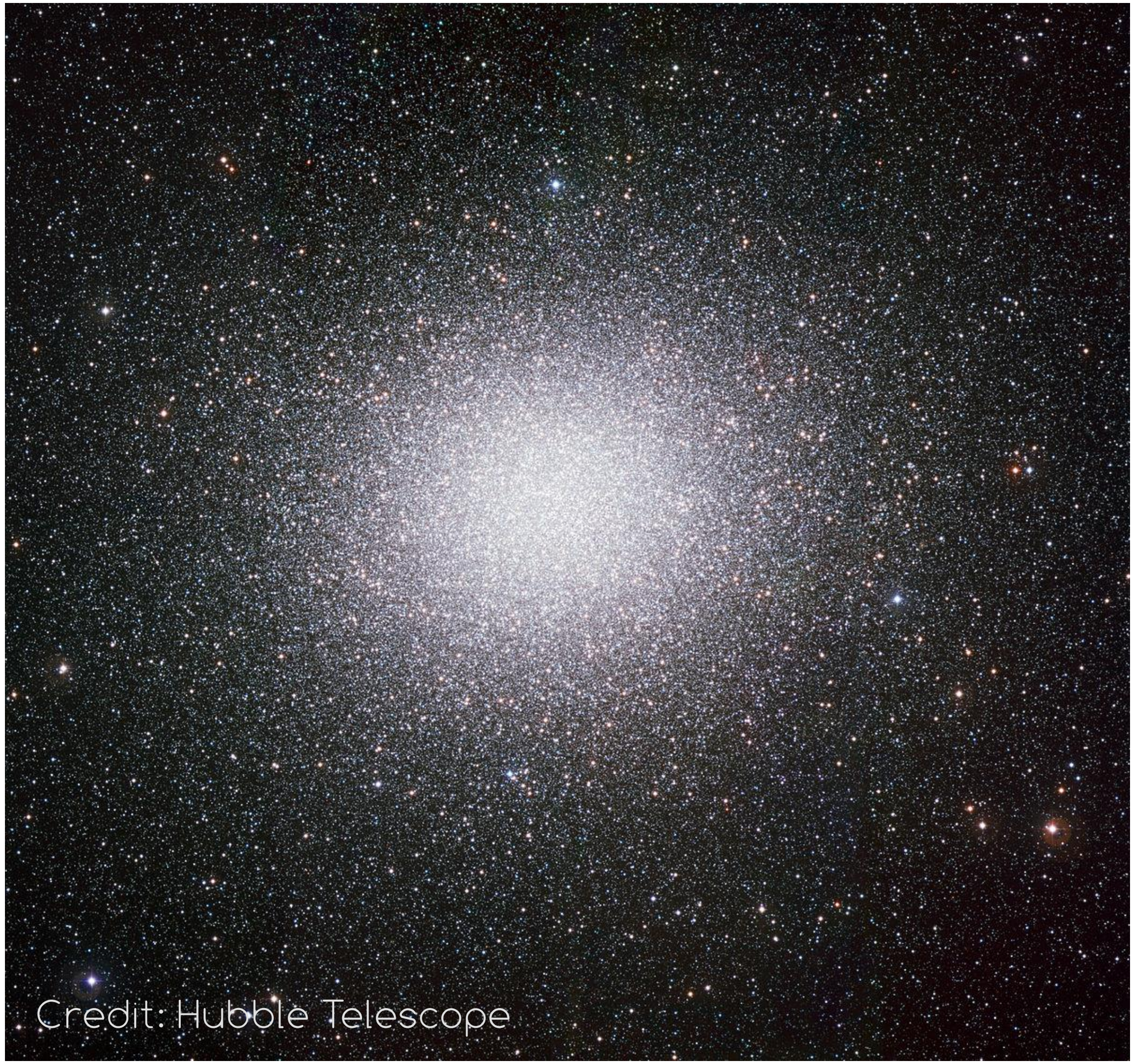
HIgh Precision PARallax COllecting Satellite

Launched 1989
Completed 1993

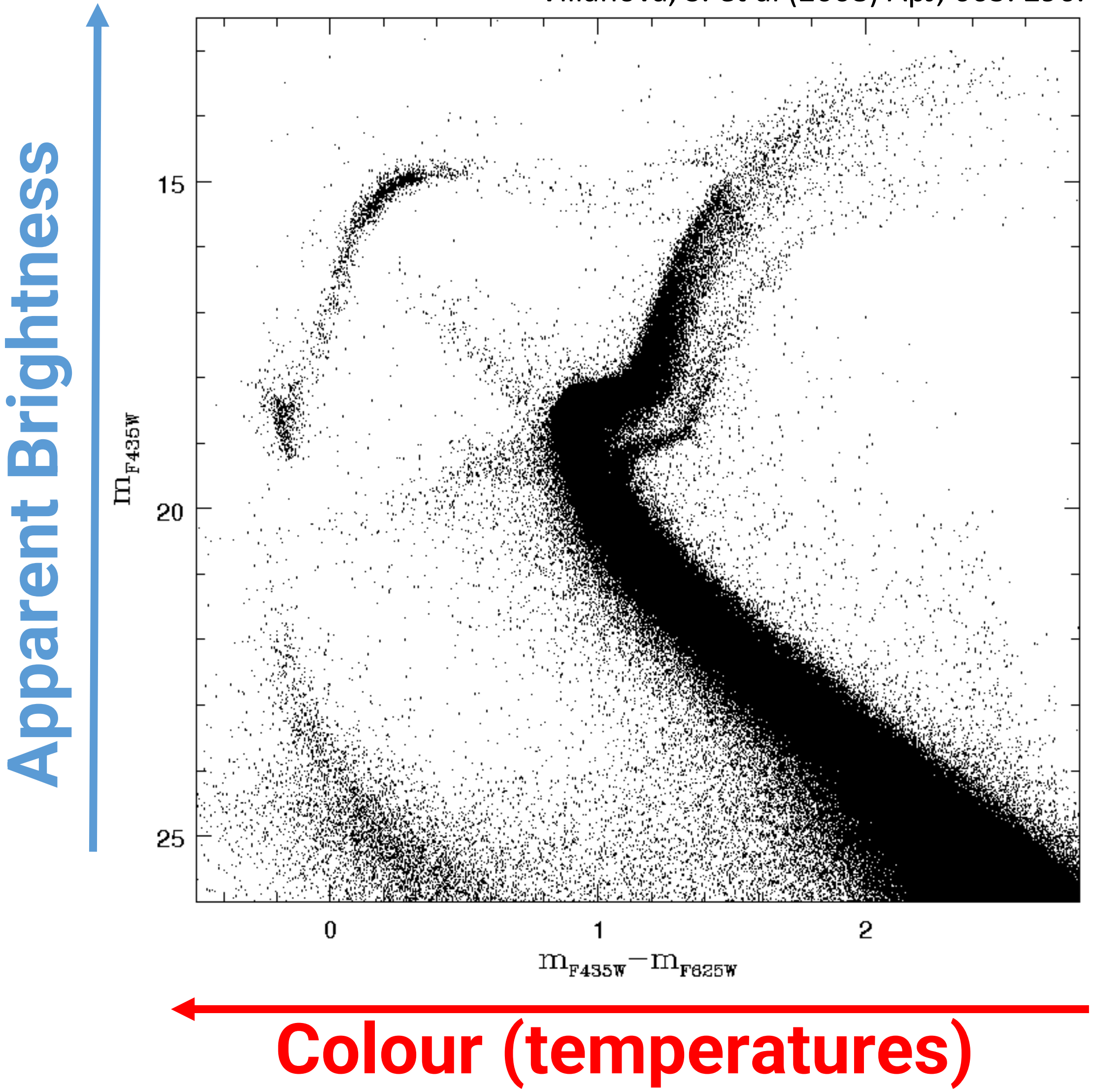
Hipparcos catalogue:

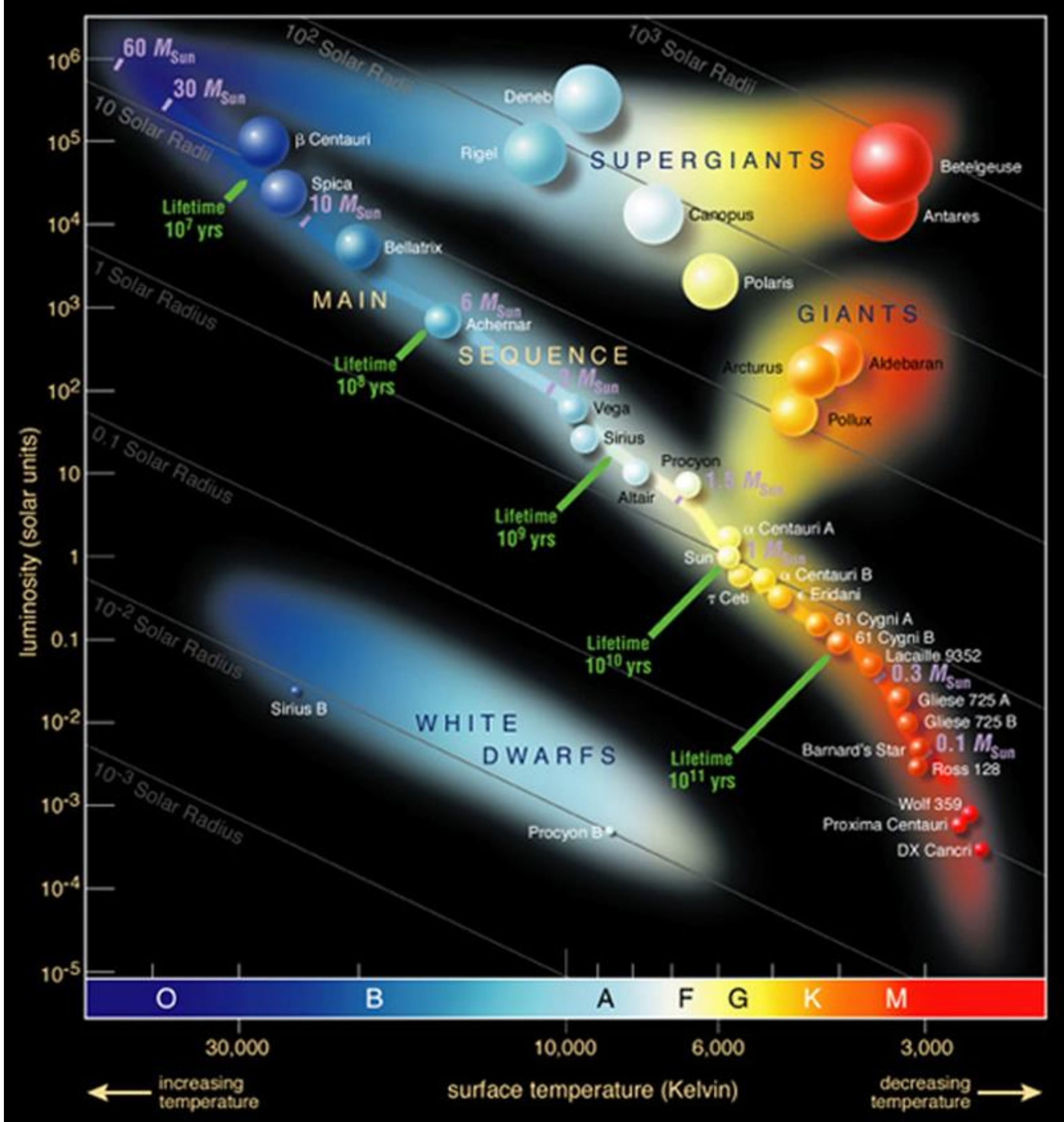
- Published 1997
- >110,000 stars
  - Position
  - Brightness
  - Proper motion
  - Parallax

Colour-Magnitude Diagrams for Globular Clusters



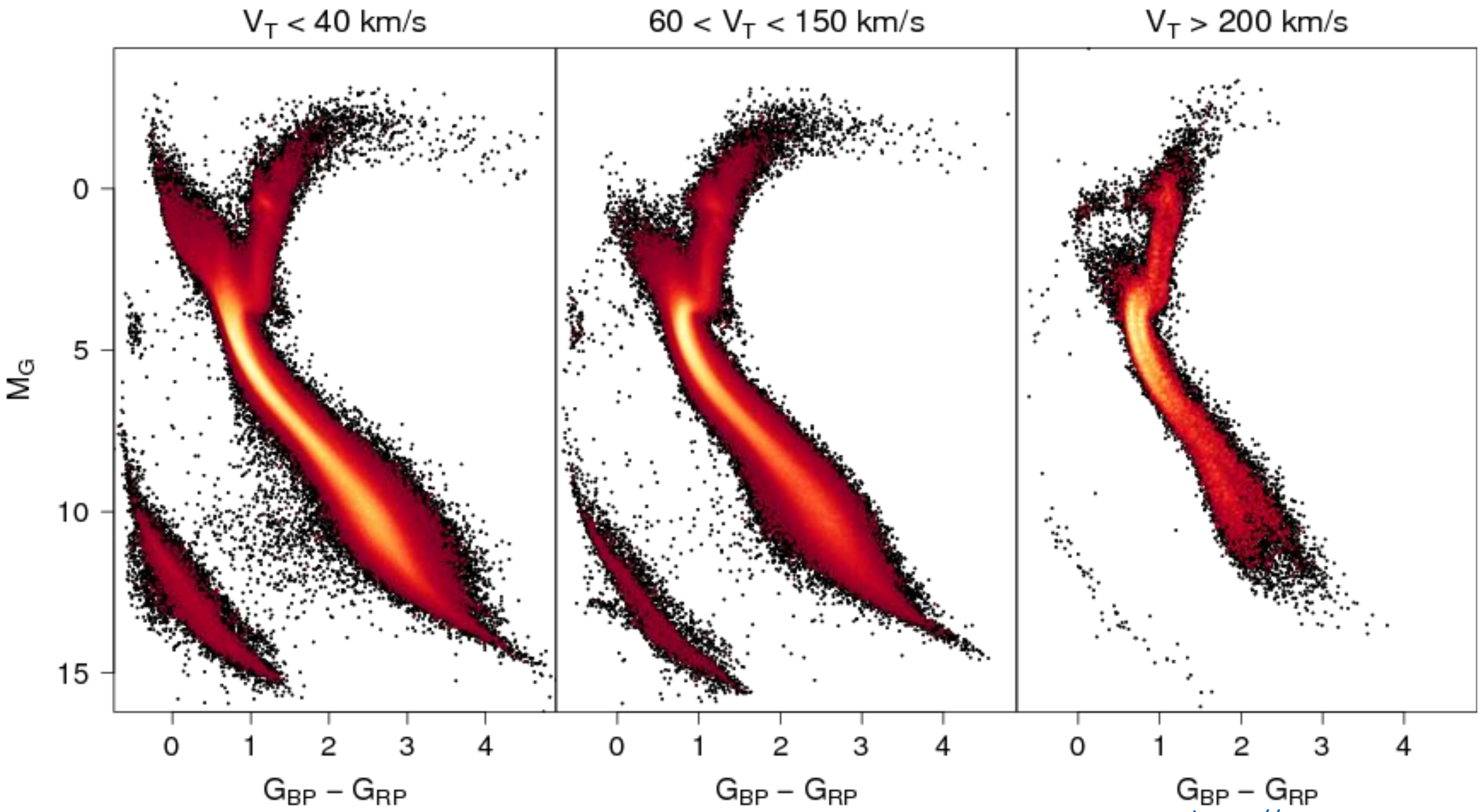
Villanova, S. et al (2008) ApJ, 663. 296.





https://www.eso.org/public/images/eso0728c/

Conditioning Plot

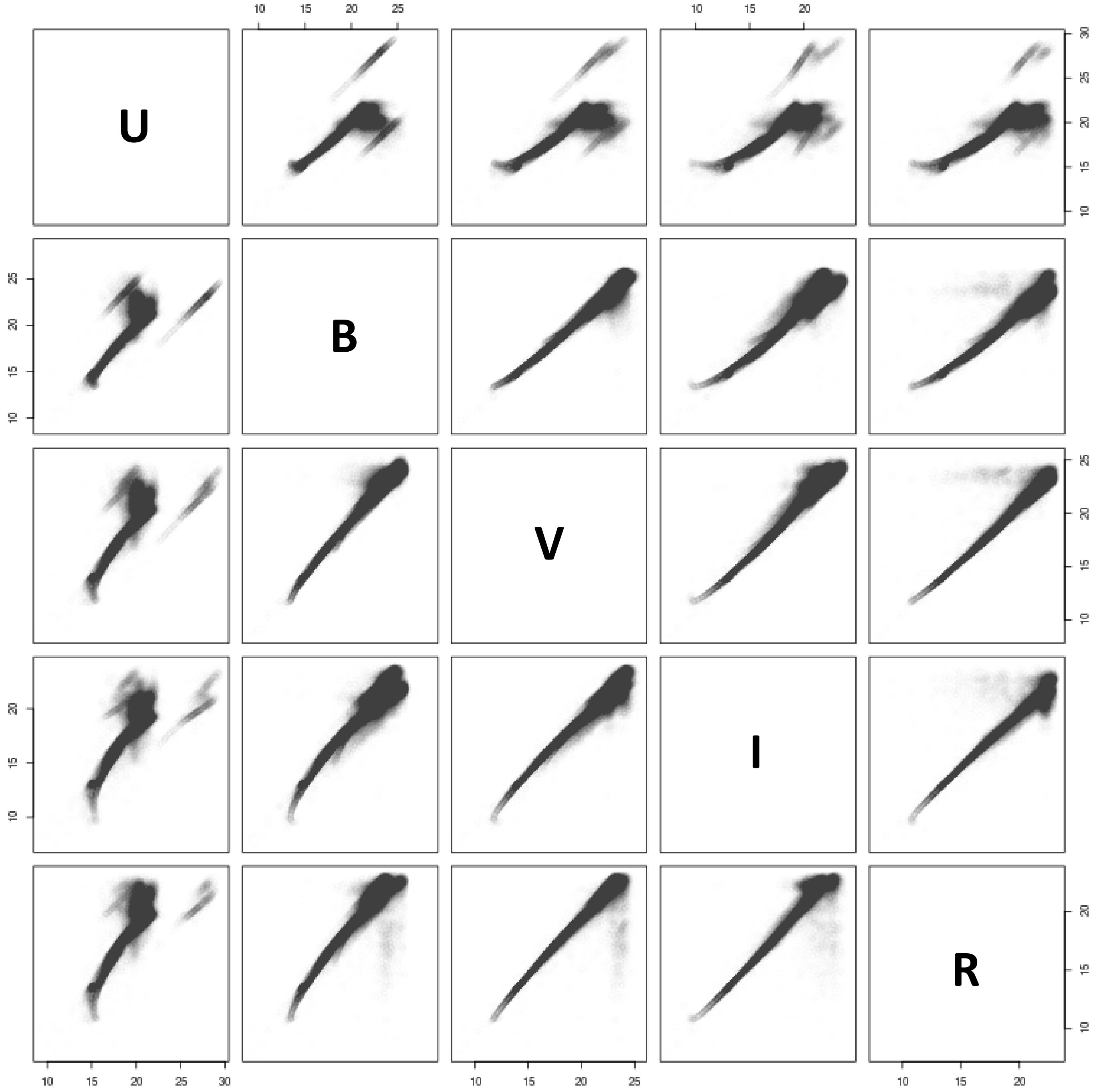


Gaia DR2

https://www.cosmos.esa.int/web/gaia/gaiadr2_hrd

Pairs Plots

- Pairs Plots are a set of scatter plots for data that have *more than two types of observations*
  - Pair all variables with all other variables



Pairs plot of U, B, V, I, R brightness measurements of stars in GC 47 Tuc

(data from Peter Stetson's catalogue)

Tukey’s 5-number summary

- Sample minimum
- Lower quartile (or first quartile)
- Sample median
- Upper quartile (or third quartile)
- Sample maximum

```
> summary(GC47tuc$B)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  9.032  19.410  21.455  21.227  23.111  28.556       3 
```

Box Plots

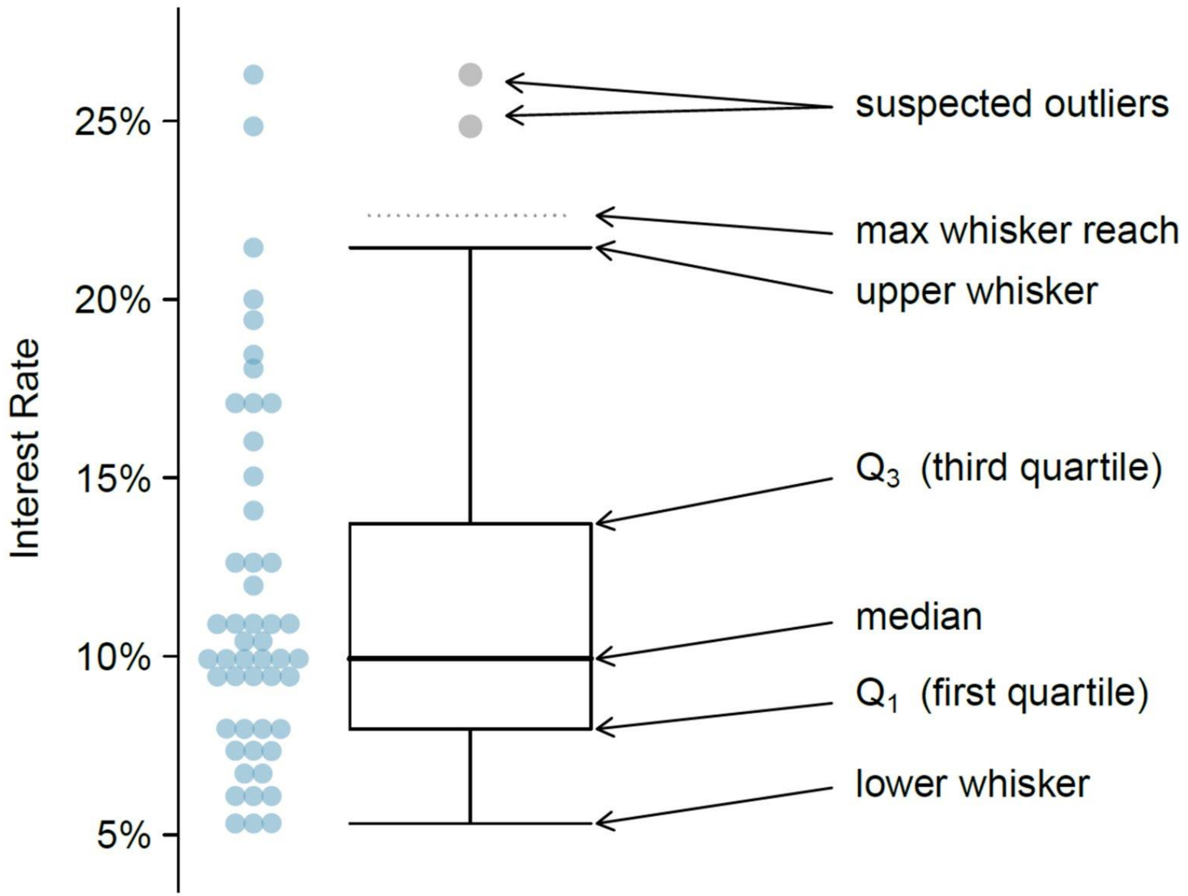


Figure 2.10, OpenIntro (4th ed.)

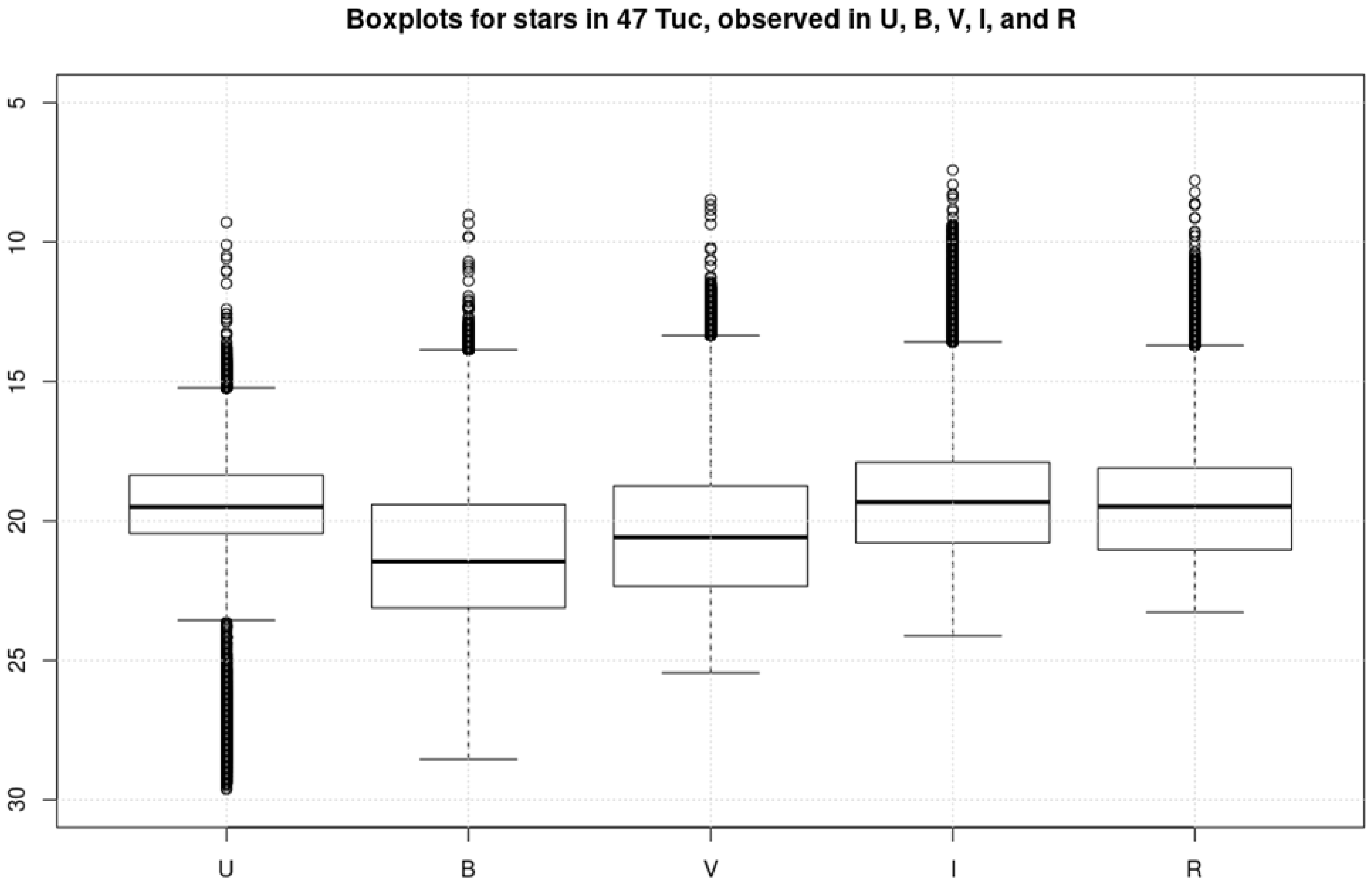
Box plot includes:

The median

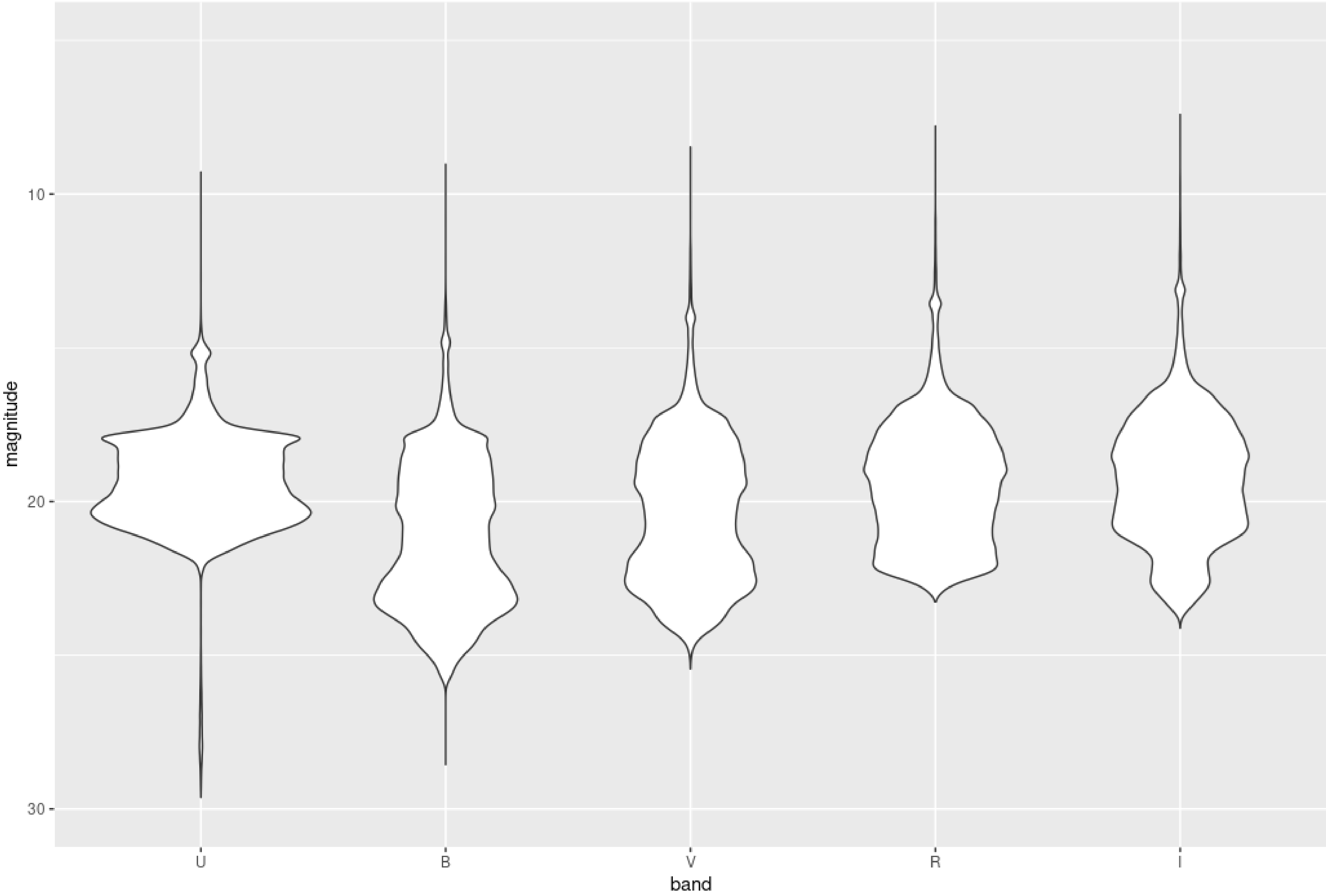
A rectangle that shows the interquartile range (IQR)

“Whiskers” out to the furthest data point that is still within 1.5xIQR

Individual points that are outside the whiskers


Boxplots for stars in 47 Tuc, observed in U, B, V, I, and R
5
10
15
20
25
30
U
B
V
I
R

Violin Plots



Mosaic Plots

Useful for categorical data

Example: Passenger data from the *Titanic*

A 4-dimensional array resulting from cross-tabulating 2201 observations on 4 variables.

No	Name	Levels
1	Class	1st, 2nd, 3rd, Crew
2	Sex	Male, Female
3	Age	Child, Adult
4	Survived	No, Yes

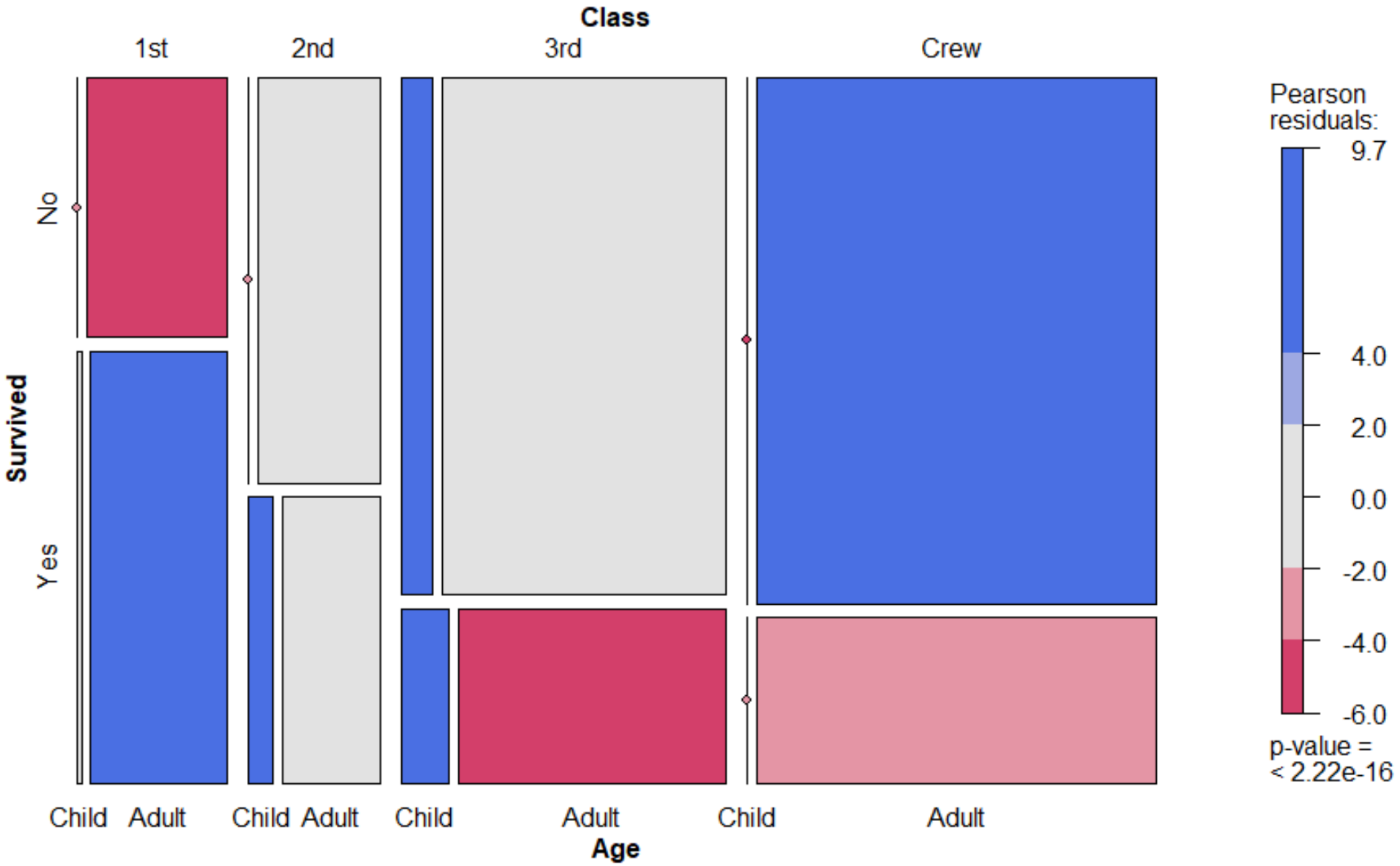
Data:
R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

```
> Titanic
, , Age = Child, Survived = No

      Sex
Class  Male Female
  1st     0      0
  2nd     0      0
  3rd    35     17
  Crew    0      0

, , Age = Adult, Survived = No

      Sex
Class  Male Female
  1st   118      4
  2nd   154     13
  3rd   387     89
  Crew  670      3

, , Age = Child, Survived = Yes

      Sex
Class  Male Female
  1st     5      1
  2nd    11     13
  3rd    13     14
  Crew    0      0

, , Age = Adult, Survived = Yes

      Sex
Class  Male Female
  1st    57    140
  2nd    14     80
  3rd    75     76
  Crew  192     20
```

Mosaic Plots

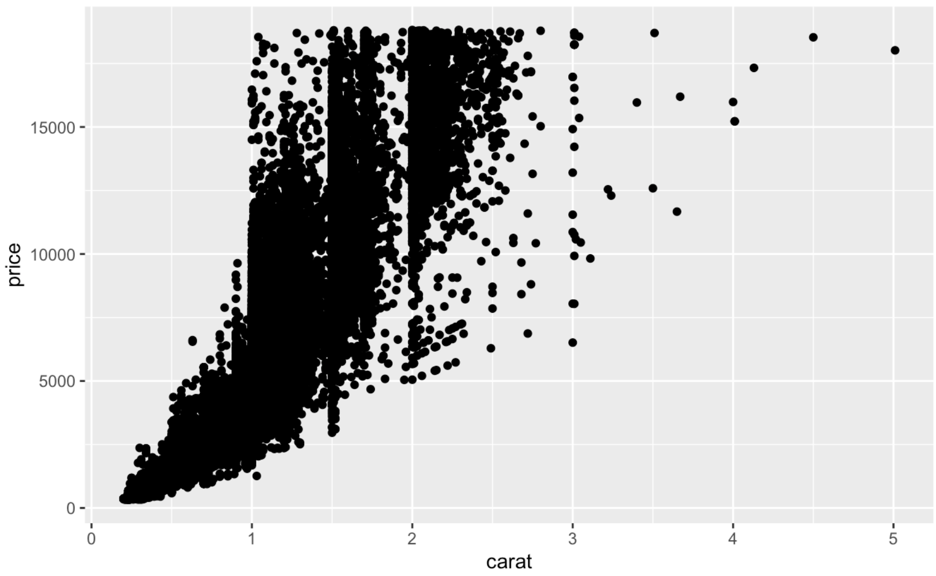
Data:
R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Code for mosaic plot:
https://stackoverflow.com/questions/40448988/mosaic-plot-and-text-values

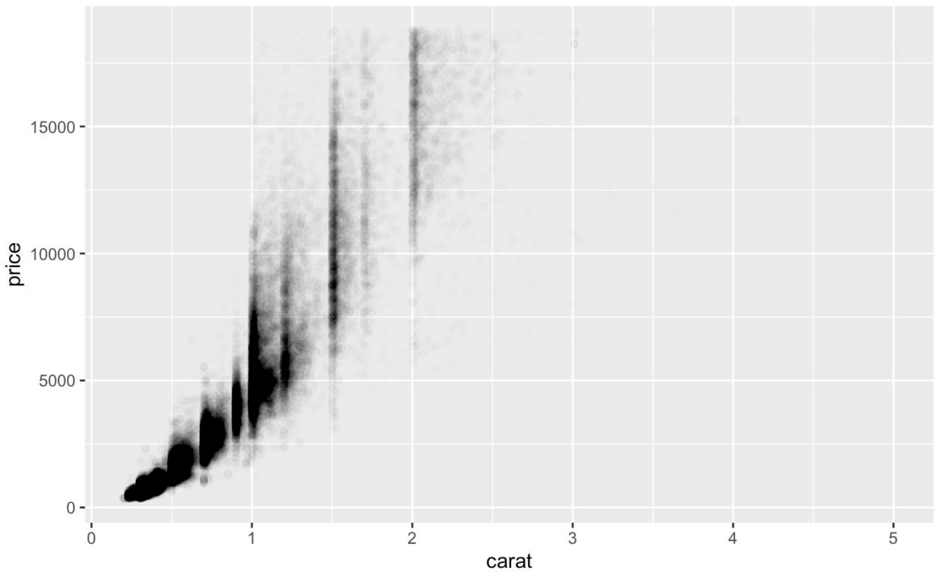
Mosaic Plots

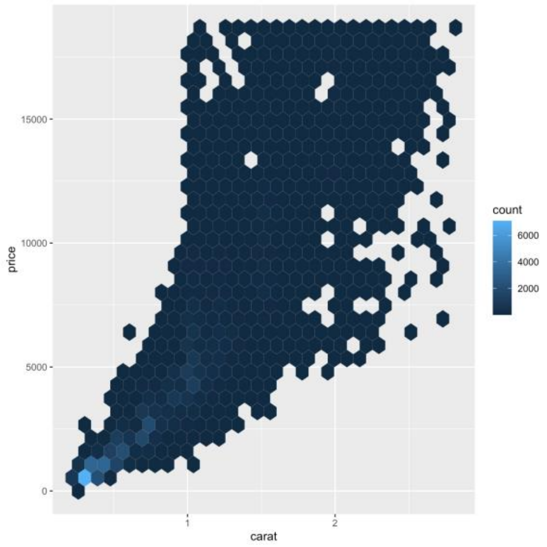


Data:
R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Code for mosaic plot:
https://stackoverflow.com/questions/40448988/mosaic-plot-and-text-values

Two continuous variables

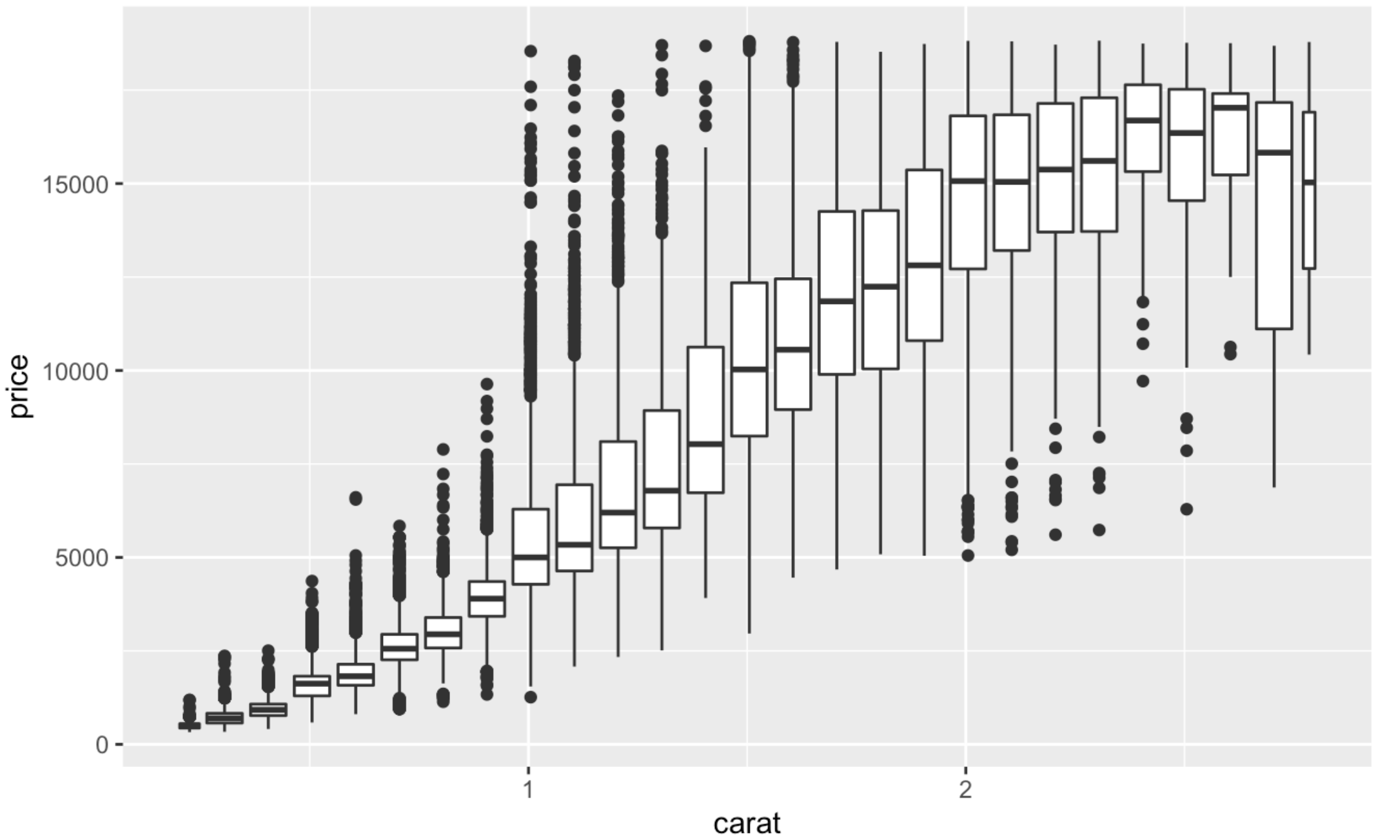






Section 7.5.3

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".



Exploratory Data Analysis and Visualization of Data

- Think carefully about how and when to plot things
  - Common mistakes:
    - putting too much information on a single figure
    - making plots when a table would be better
    - Making a table when a plot would be better
    - Only using one EDA tool from the EDA toolbox
- Humans are easily fooled when comparing areas, volumes, colours, curvature, etc.

Graphical Perception

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

Source: *Journal of the American Statistical Association*, Sep., 1984, Vol. 79, No. 387 (Sep., 1984), pp. 531-554

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2288400

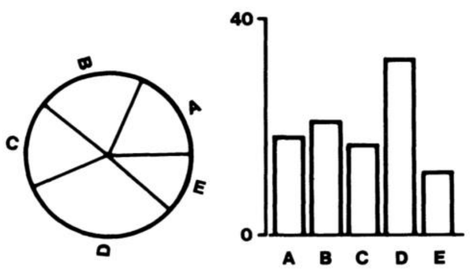


Figure 3. Graphs from position–angle experiment.

The following are the 10 elementary tasks in Figure 1, ordered from most to least accurate:

1. Position along a common scale
2. Positions along nonaligned scales
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color saturation

Journal of the American Statistical Association, September 1984

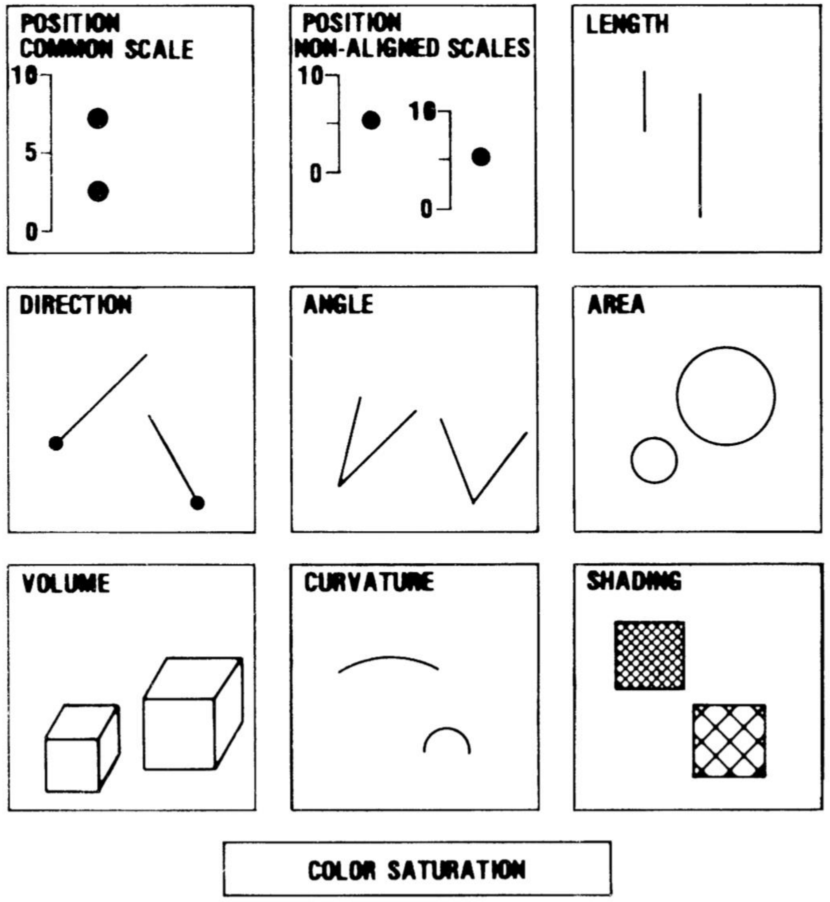
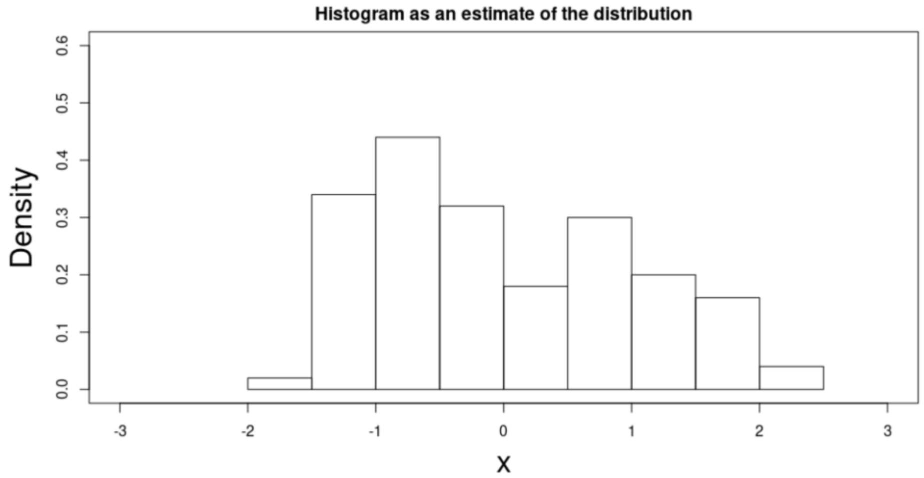


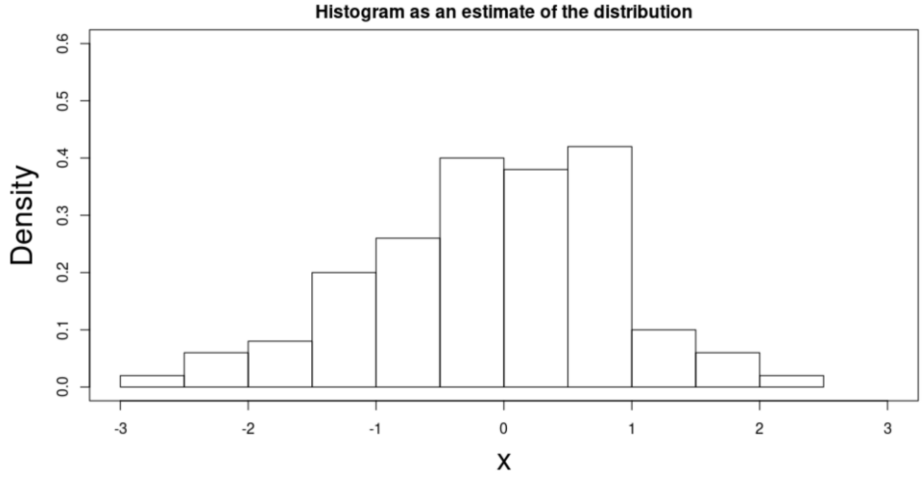
Figure 1. Elementary perceptual tasks.

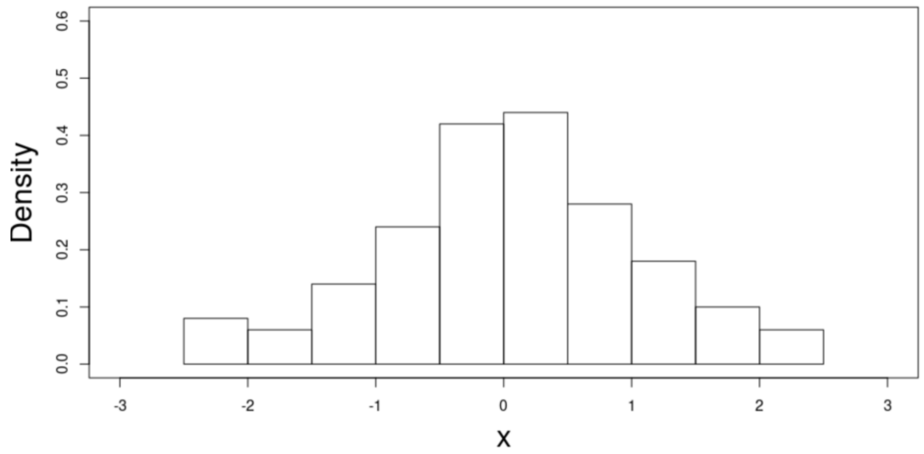
Randomness in Data

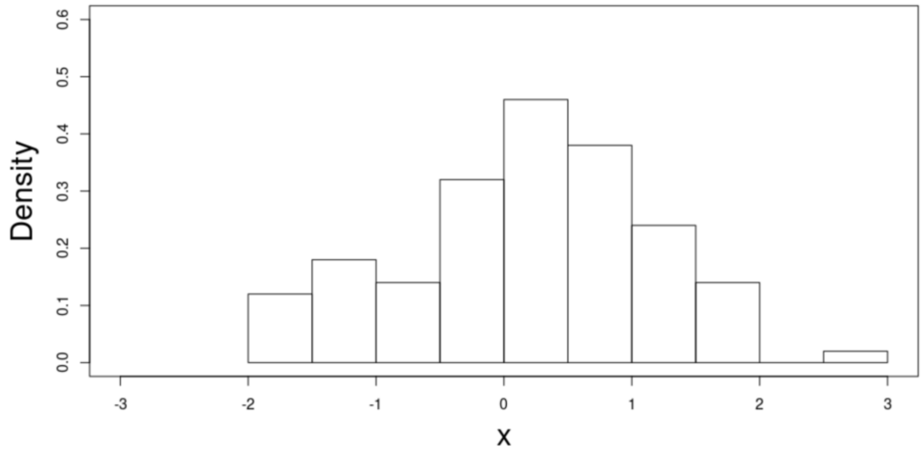
- All these histograms were generated from 100 draws from a standard normal

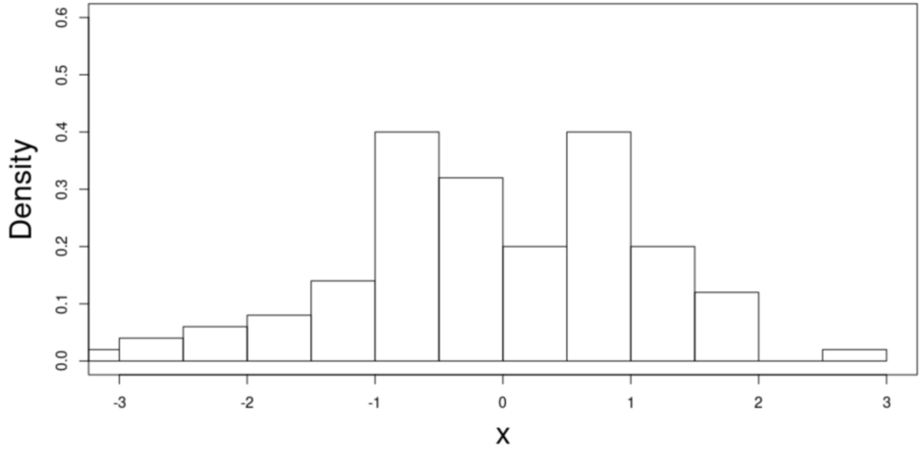


From the data, we can try to estimate the parameters of the underlying distribution.









EMPIRICALLY SUMMARIZING DATA

Mean

- The mean is the average
- *The sample mean is:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The mean is not a robust statistic, because it is not "resistant" to extreme observations

- The true or population mean is usually denoted \mu, and is often unknown

Sample Variance

- The average squared distance from the mean
- *The sample variance is:*

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

- What does the squaring do?
  - Makes all differences positive
  - Makes large differences relatively much larger

- The true or population variance is usually denoted \sigma^2, and is often unknown

Variance

- *Distributions don't have to look the same to have the same variance*

- What other ways could you describe these distributions to differentiate them?

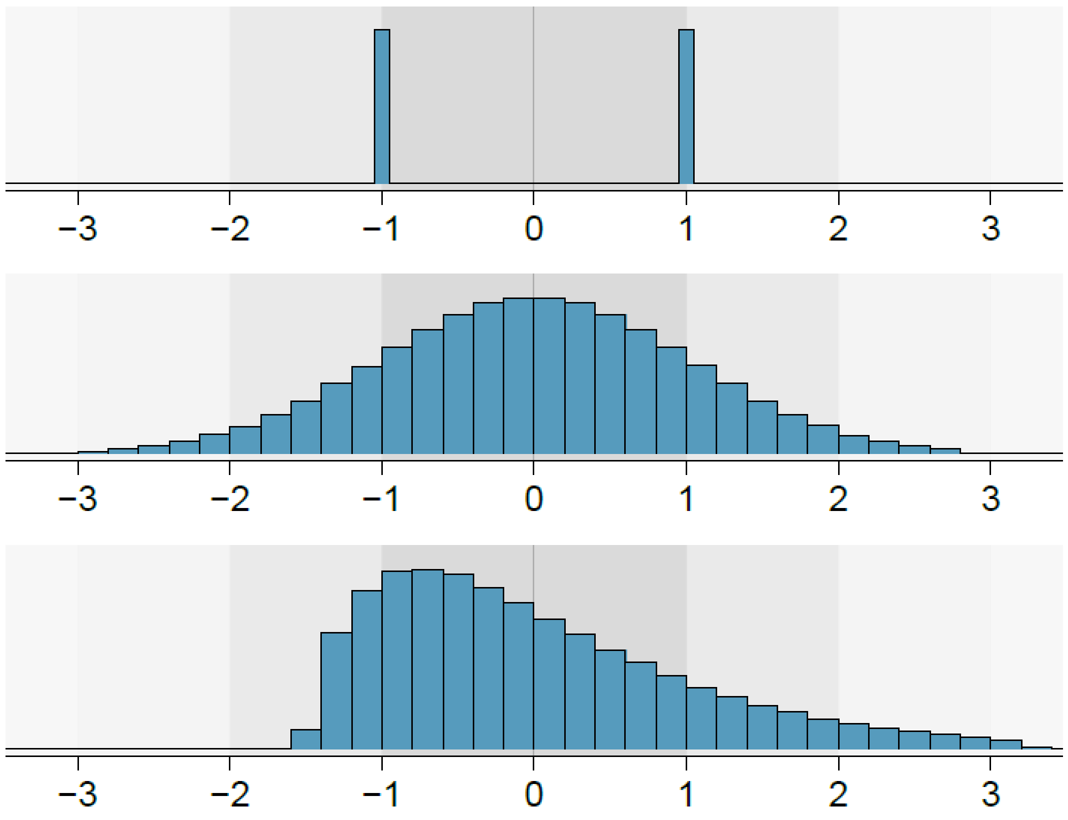


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

PERFORMING INFERENCE WITHIN A BAYESIAN FRAMEWORK

- Requires understanding
  - random variables and how they relate to data
  - some basic probability rules
  - common parametric distributions
  - How to write down a likelihood
  - *Philosophical difference in interpretation compared to frequentist framework*

RANDOM VARIABLES

Random Variables

Assigns a *numerical* value to an outcome/event from an experiment

- Notation:

$$X, Y, W, \ldots etc.$$

- Examples of random variables:
  - Birth weight of a baby
  - Height of a person → continuous random variable
  - How long you wait for the bus
  - The winning score for a basketball game → discrete random variable

Random Variables

Can be *continuous* or *discrete*

- What are some examples of *continuous random variables* in astronomy?
- What are some examples of *discrete random variables* in astronomy?

A (data) *sample* is a realization of a random variable

- Imagine we measured the brightnesses of 25 randomly selected stars from the sky

  →These data are realizations x of a random variable X that represents the brightness of a star

- To perform (parametric) *statistical inference* on our data x, we assume how the random variable X is distributed

- Once we have a statistical model for how X is distributed, then we can say things like *"the probability of observing a star with magnitude less than 17 is …"*

Randomness matters!

- We usually don't know how X is distributed!
- Our *random* data sample x is just that – a sample!
- Randomness can trick us! Humans like to look for patterns
- Things get even trickier when our data samples suffer from selection bias, observation bias, etc.
- We should look at and summarize our data in different ways. Be skeptical.
  - → Exploratory Data Analysis

Modelling data as a random variable

Standard statistics notation to show what distribution a random variable follows:

E.g., we might assume that data x (e.g. the photon counts from a star) follows a Poisson distribution:

DISTRIBUTIONS

A DISTRIBUTION…

Tells you the frequency or relative frequency of each possible value/event, or of some data that was collected

Could be empirical or analytic

Can be useful for modelling a population of objects

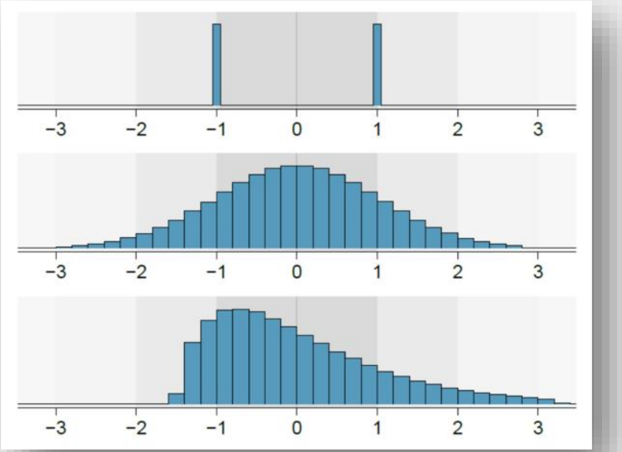
Is often a foundation of statistical reasoning

Can be continuous or discrete

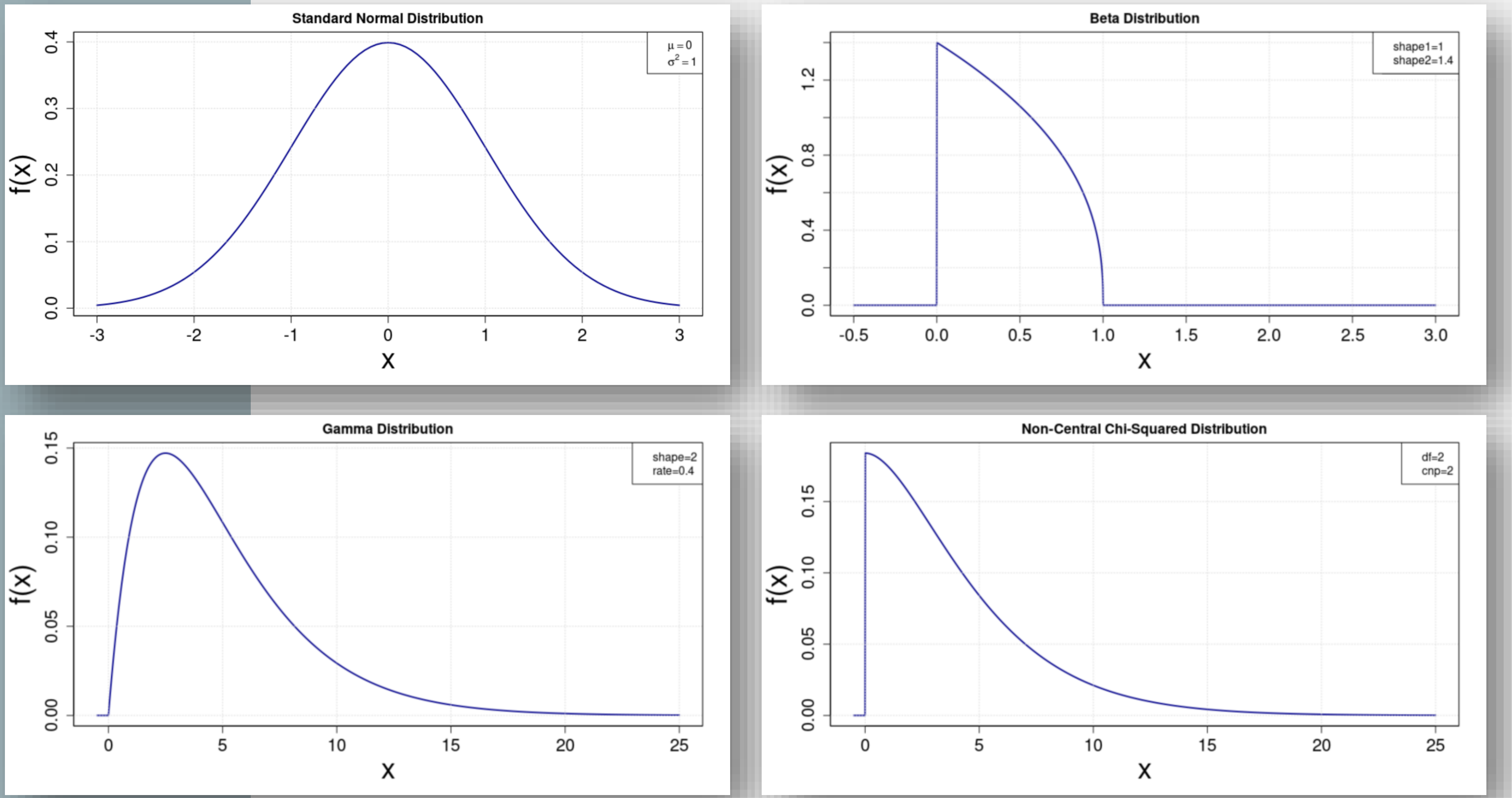
That is analytic has parameters that define its shape

Can be univariate or multivariate

Example histograms (figure from Open Intro Statistics 4th ed.)



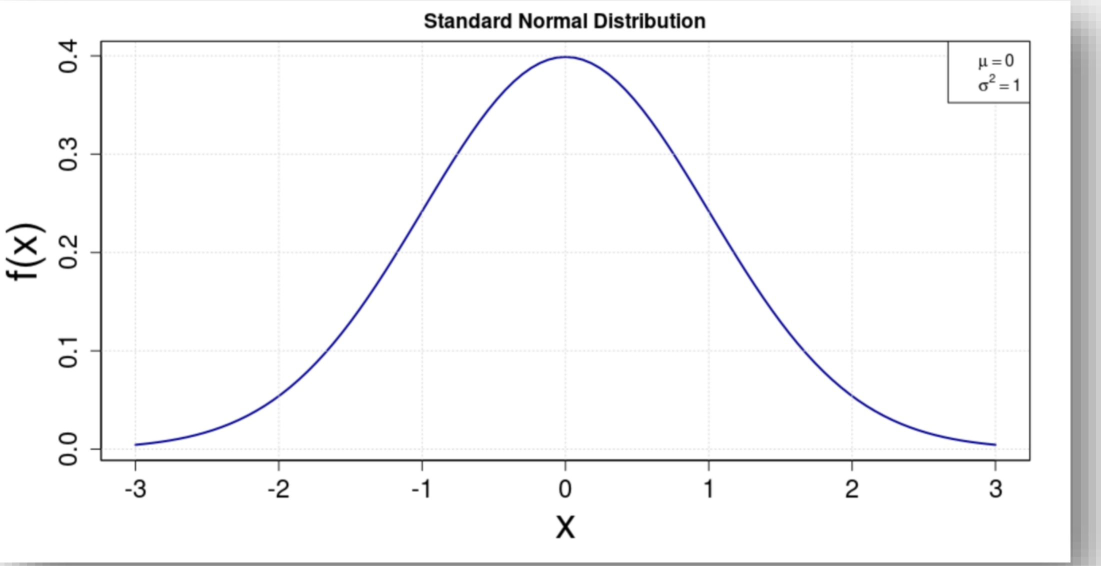
Some analytic probability distributions



PROBABILITY DISTRIBUTIONS

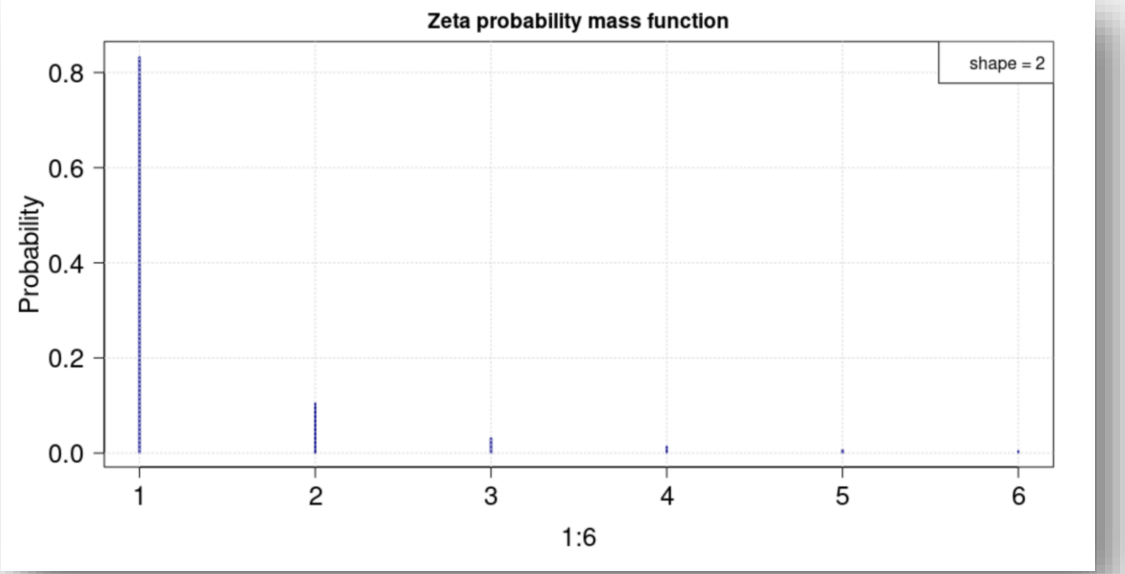
Continuous quantities

probability density function (pdf)



Discrete quantities

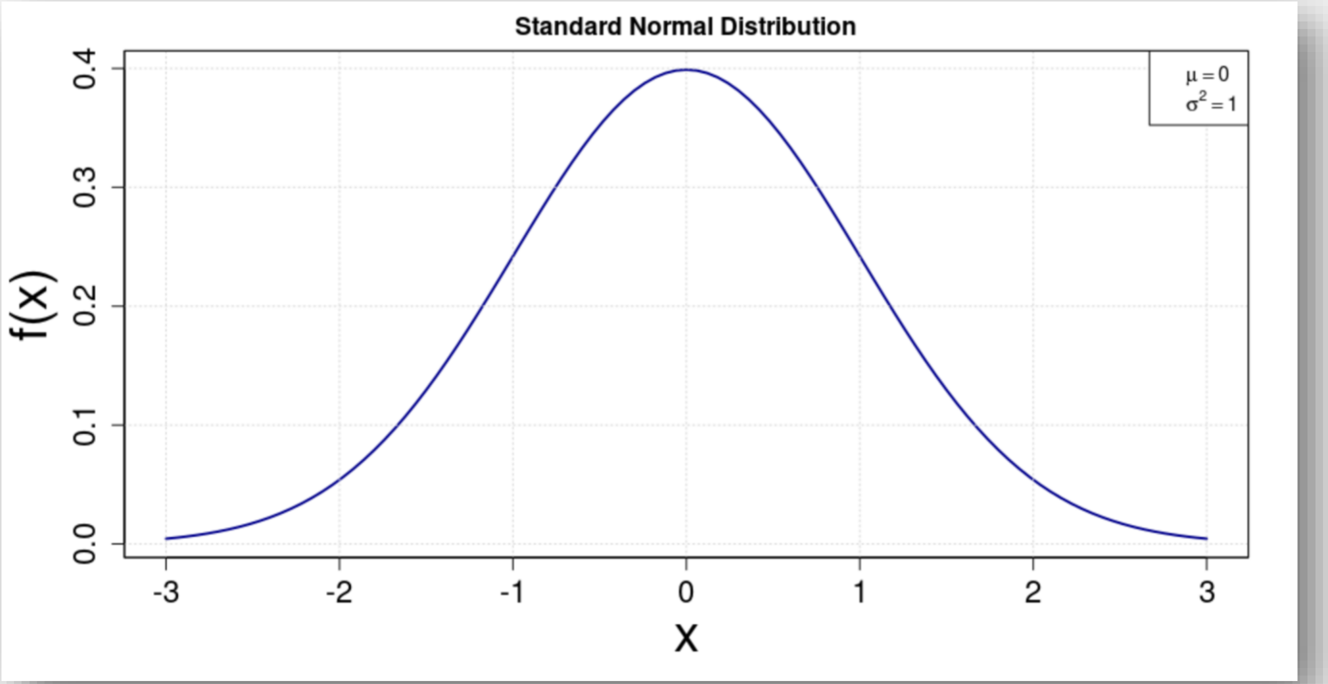
Probability mass function (pmf)



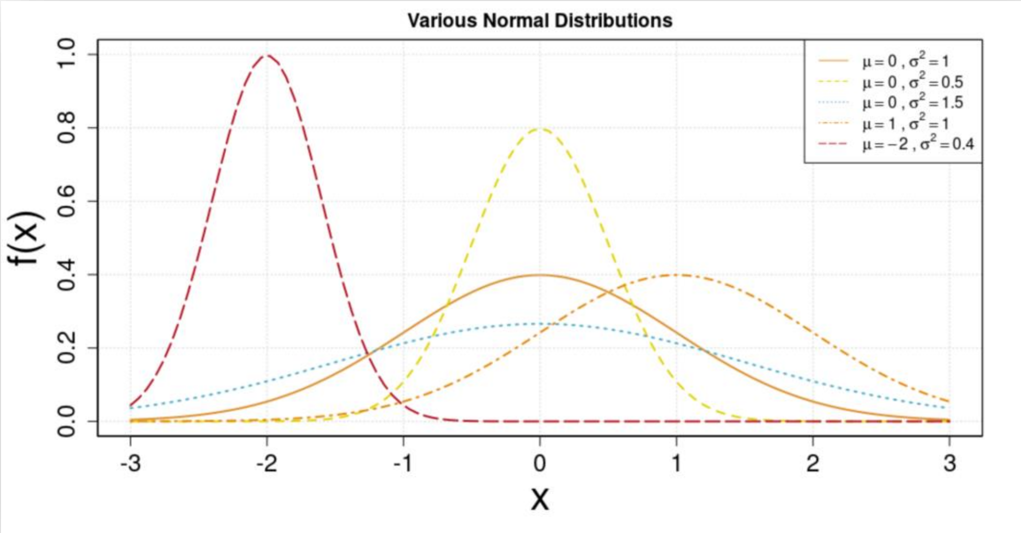
THE NORMAL DISTRIBUTION

$$N(\mu, \sigma^2)$$

The mean and variance entirely define the Normal → if you know these you can plot it



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



Empirical Rule – "68-95-99 rule"

Normal or Gaussian distribution is defined by a **mean** and a **variance.**

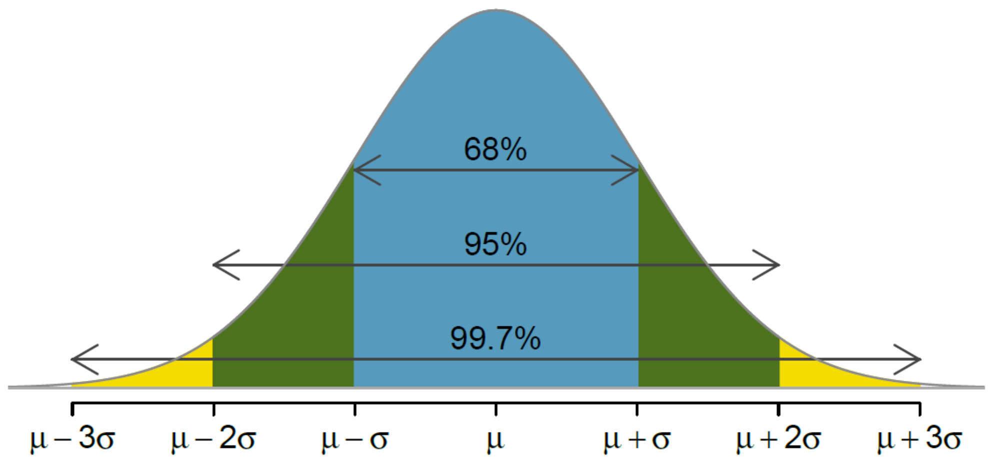


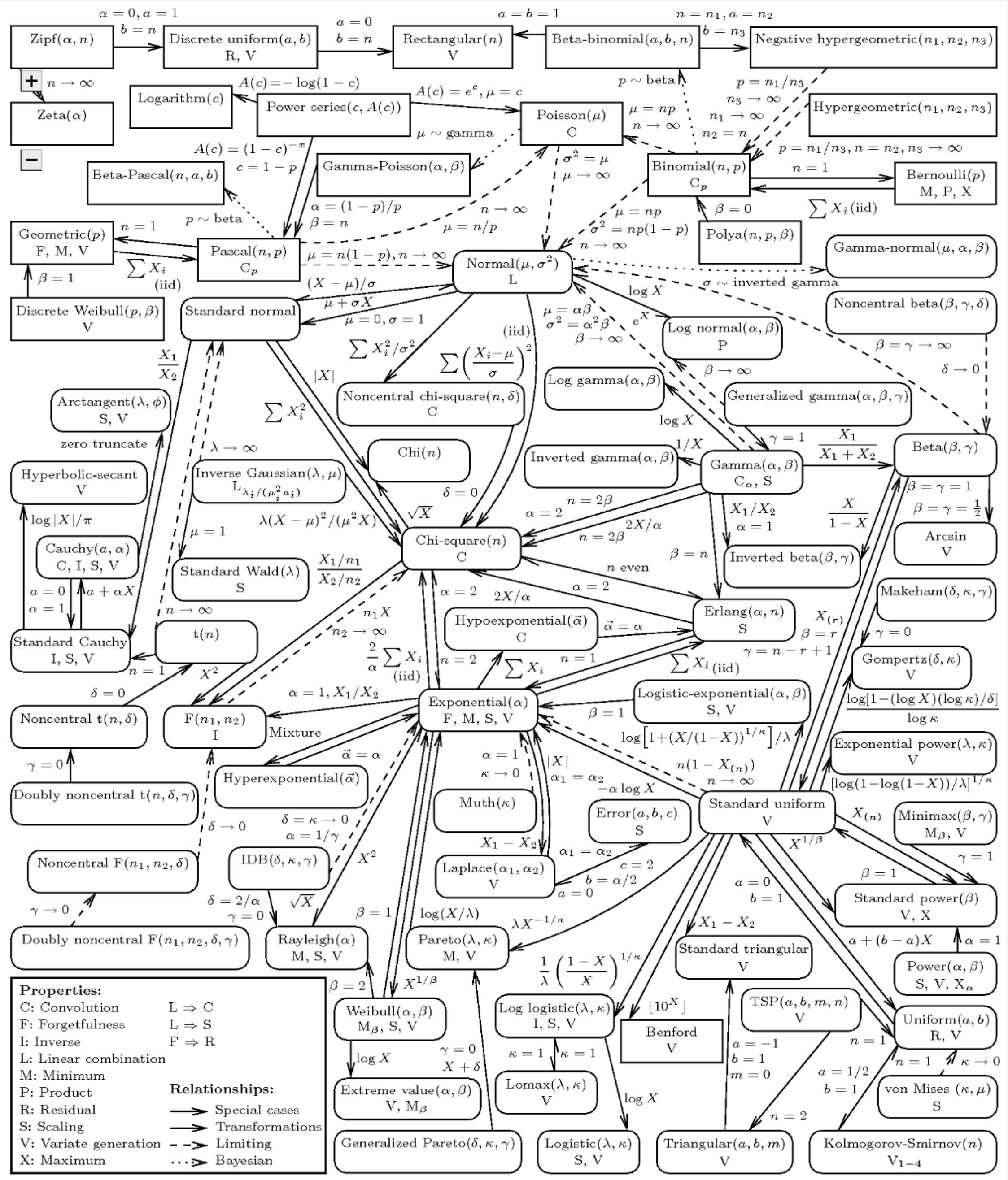
Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

IMPORTANT NOTE:

68% is equivalent to 1σ (1 standard deviation) only in the case of Normal/Gaussian distributions!

The 68% quantile is not necessarily equal to 1σ in non-Gaussian distributions

THERE ARE MANY UNIVARIATE DISTRIBUTIONS!



http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

JOINT, CONDITIONAL, AND MARGINAL DISTRIBUTIONS

Joint, Conditional, and Marginal Distributions

Joint Distribution

- The 2-d (or more!) distribution of two or more things

Conditional Distribution

- The distribution of a variable given an event or value for another variable

Marginal Distribution

- The distribution of a variable *regardless* of the values of the other variables