Statistics for CSAI II

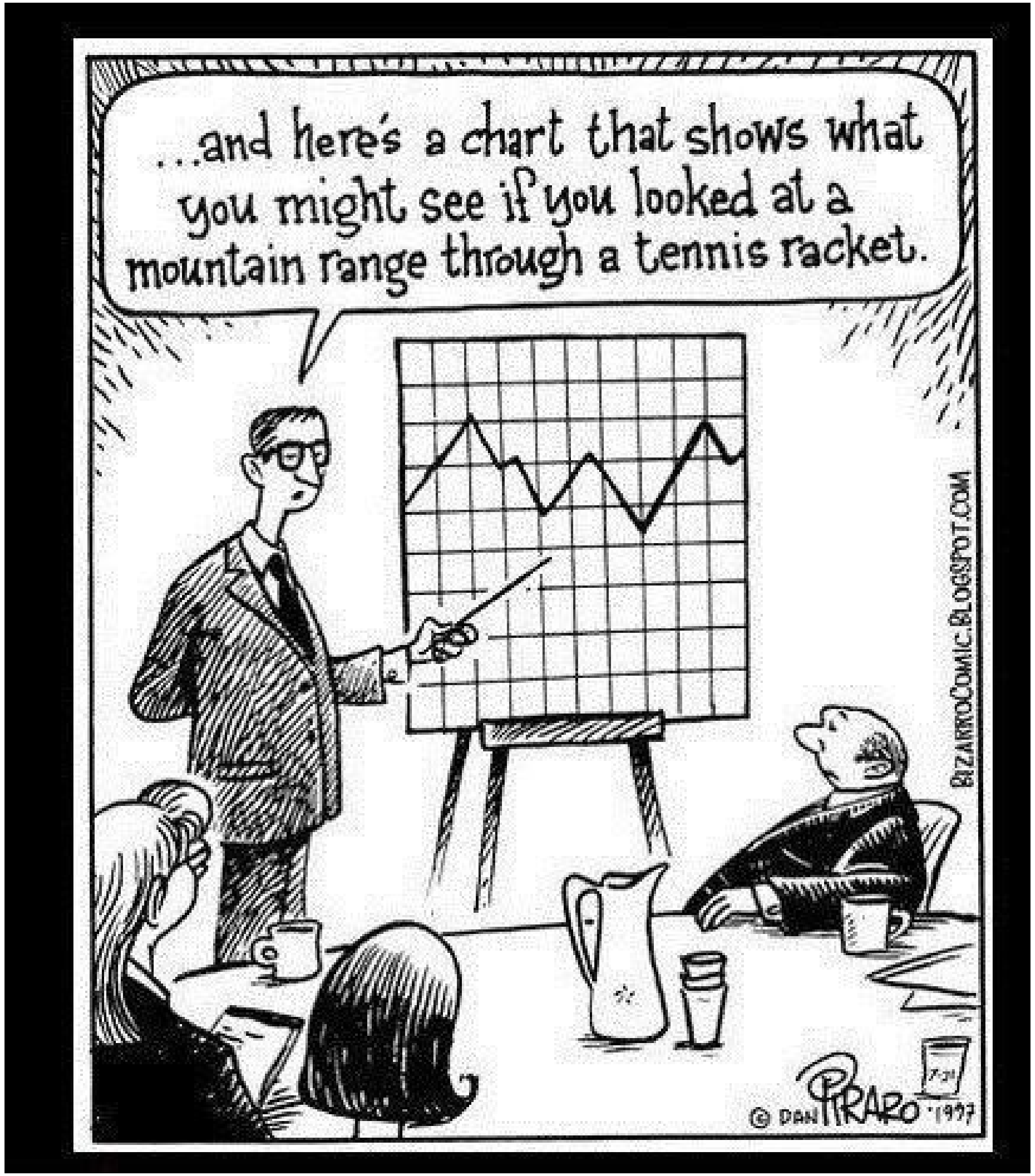
Introduction

Dr. Travis J. Wiltshire

TILBURG UNIVERSITY

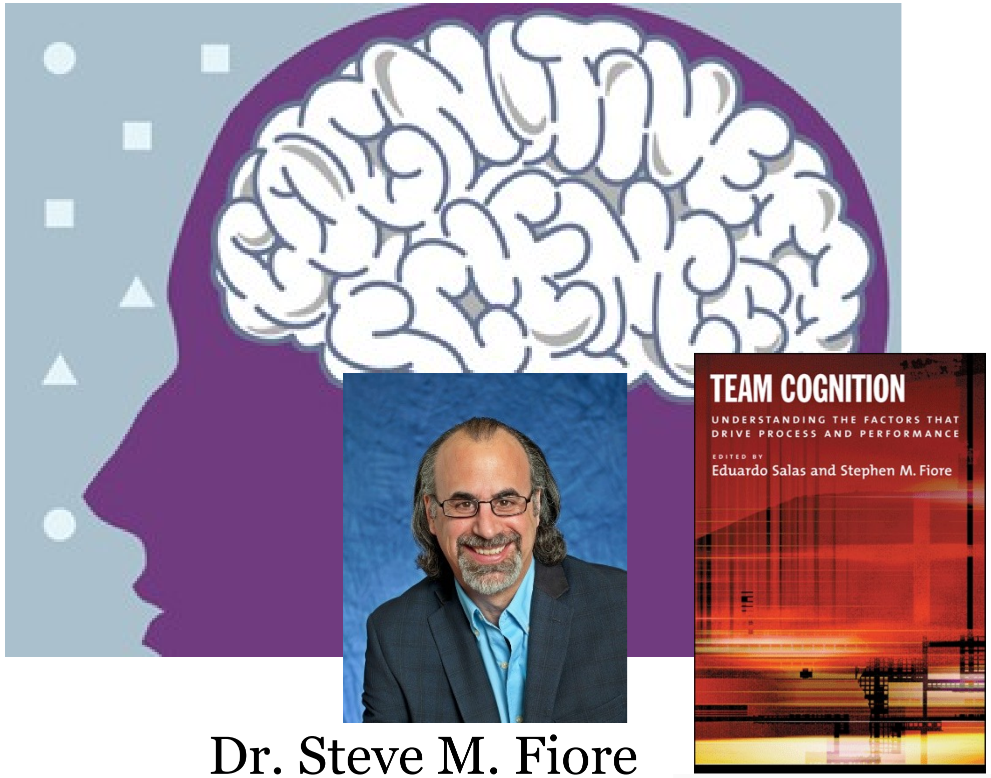
Outline

1. Course Expectations
2. R, Rstudio
3. Probability Basics


...and here's a chart that shows what you might see if you looked at a mountain range through a tennis racket.
BIZARROCOMIC.BLOGSPOT.COM
© DAN PIRARO 1997
7-21

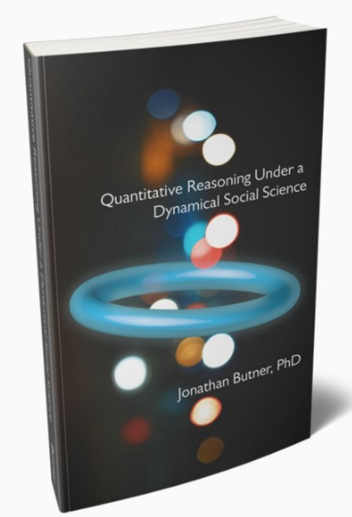
About Me

- B.S. in **Psychology**, University of Central Florida, USA
- M.S. in **Human Factors & Systems**, Embry-Riddle Aeronautical University, USA
- PGC in **Cognitive Science**, University of Central Florida, USA
- Ph.D. in **Modeling & Simulation**, University of Central Florida, USA
- Postdoctoral Fellowship, **Quantitative Psychology and Dynamical Systems**, University of Utah, USA
- Postdoctoral Fellow, Carlsberg Foundation, Department of Language and Communication, SDU, DK


COGNITIVE SCIENCE
TEAM COGNITION
UNDERSTANDING THE FACTORS THAT DRIVE PROCESS AND PERFORMANCE
EDITED BY
Eduardo Salas and Stephen M. Fiore

Dr. Steve M. Fiore

Dr. Jonathan Butner



CARLSBERG FOUNDATION

Our Teaching Team

Sasha Kenjeeva
Co-lecturer (Practical sessions)

PhD student dept. CSAI as of September 1st, 2023

Barbora Lukáčová
Teaching Assistant

MSc student CSAI

Quiz

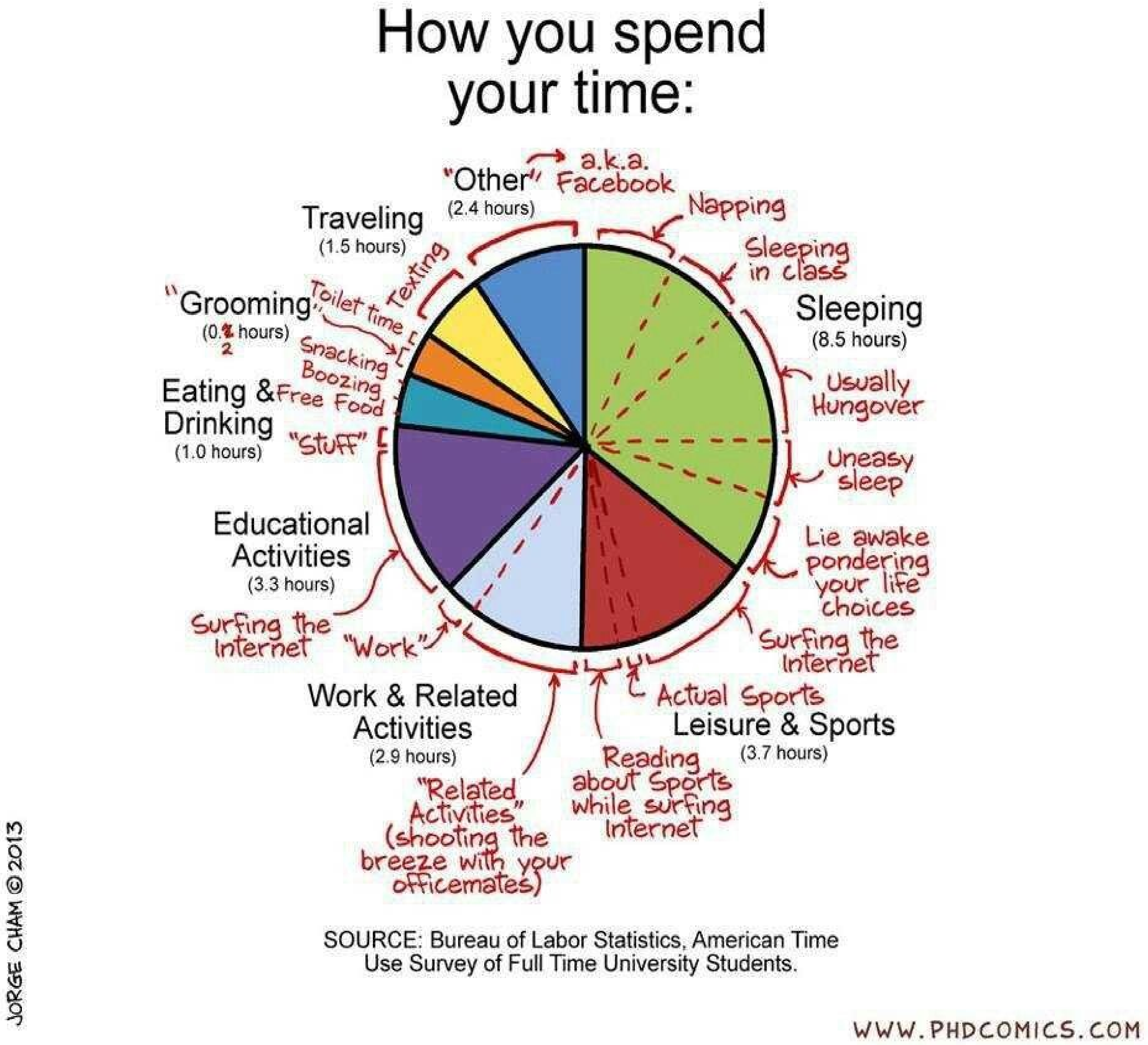
Course Expectations

- Read the Syllabus!
- Follow the course schedule.

Course Expectations

- 6 EC course:
  - 28 hours per EC

Total hours for course	**168**
Time in Lecture	21
Time in Practical Sessions	10
Time for reading	45
Time for practicing in R	40
Time for Completing Practical Assignments	22
Time to prepare for Exams	30



SOURCE: Bureau of Labor Statistics, American Time Use Survey of Full Time University Students.

How to get the most from this course (and be good at statistics in the future):

- To get the most out of this course, I recommend for each topic/module you:
  - Read the Required Text
  - Complete Swirl/LearnR Tutorials
  - Complete Pre-Class Quizzes
  - Attend/Watch Lectures
  - Engage in in Class Exercises
  - Check your Understanding (Reread if necessary, post questions on Canvas)
  - Complete Required Practical Exercises
  - Evaluate Practical Exercise Solutions (Post questions on Canvas, ask Questions in Practical Sessions)
  - (Optional) Complete Associated Open Stats Lab)

- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public *Interest, 14*(1), 4-58.

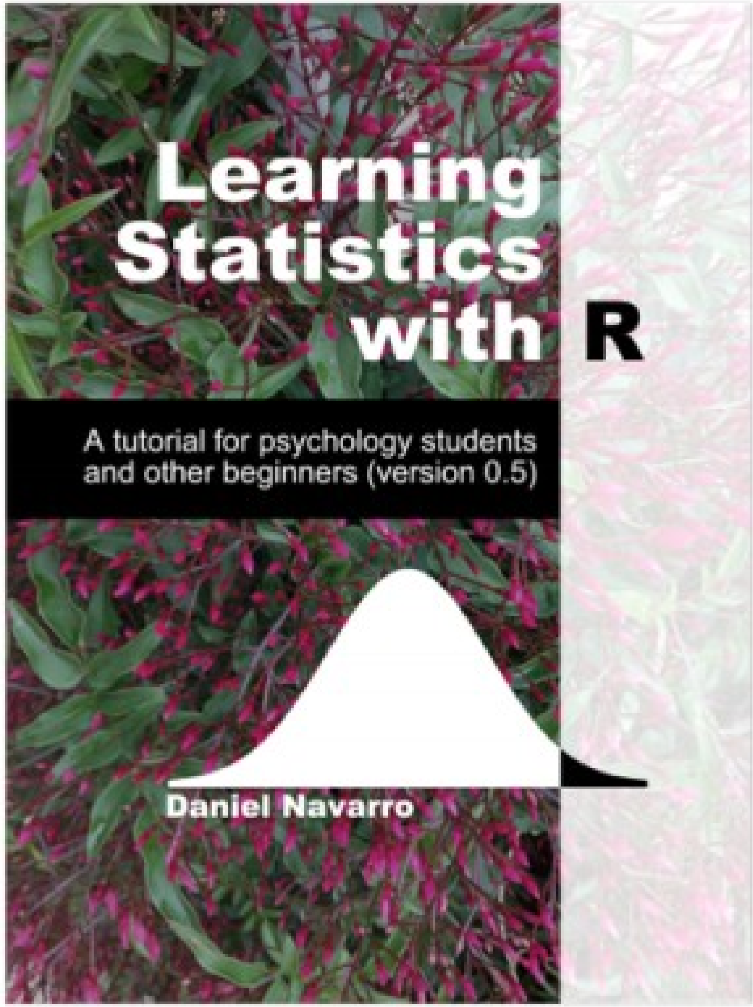
Course Expectations

- Our Textbook is free!
  - Download it from here:

https://compcogscisydney.org/learning-statistics-with-r/

Other reading materials are uploaded on Canvas.

Required readings must be completed before class on the date listed in the schedule!

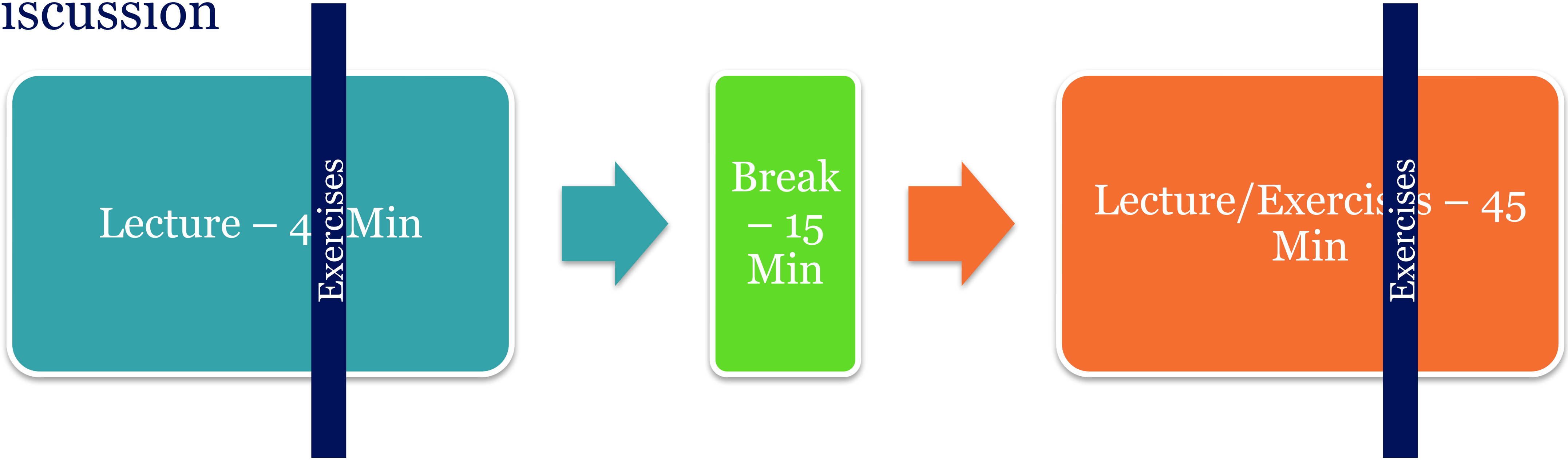

Learning Statistics with R
A tutorial for psychology students and other beginners (version 0.5)
Daniel Navarro

Course Learning Objectives

- Understand basic probability and sampling theory and how that informs our understanding of statistical analyses and null hypothesis significance testing
- Conduct and interpret associations between variables with correlations using R
- Conduct and interpret output from linear regression models using R
- Conduct and interpret output from linear mixed-effect models using R

Lecture and Practicals

- Lecture meeting time and location: See My Timetable, location and time can vary.
- Practical Sessions: Typically on Fridays (see My Timetable). Currently two time slots available. Tutorials / Q&A / Exercise preparation and discussion



Lecture – 45 Min

Exercises

Break – 15 Min

Lecture/Exercises – 45 Min

Exercises

Course Expectations

- Canvas
  - All course information and materials will be posted here
  - Submitting Practical Exercises
  - Post/reply to questions on the discussion forum
  - Questions should be sent through Canvas.

Course Expectations

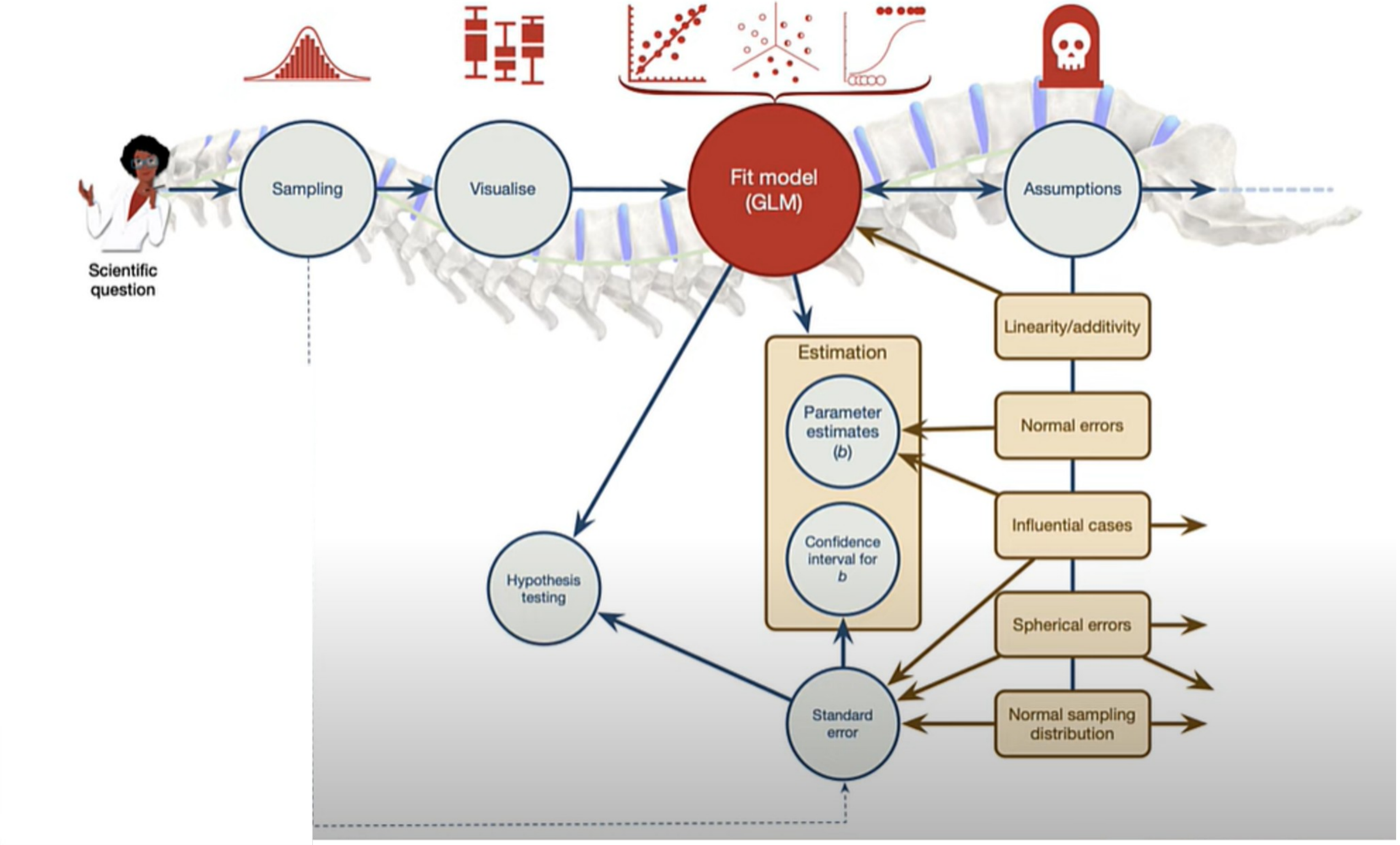
- **Exams (80% of your grade)**
  - Final Exam – 80%
  - Multiple choice + Constructed Response (e.g., Fill in the Blank)
  - **Date and times for exams and resits will be posted in your time table**
  - **You must make sure to register for the exams!**

- **Practical Sessions/Exercises (20% of your grade)**
  - 11 in person practical sessions (see TimeEdit) where we make exercises available **during the sessions with a password**
  - 10 assignments submitted through Canvas **due at the end of an in person practical session**
  - **No resubmissions or resits** (leniency policy: (8/10 required for full points, no skipping last two)
  - Will be checked for completeness
  - Self-evaluation- You check the solutions posted/discuss in Practical Session
- **Recommended Exercises/Tutorials (in LearnR)**
  - Listed in schedule

Course Expectations

- ***Questions about statistical theory or analyses:***
  - Check the syllabus, the required readings, or in the lecture slides.
  - Post your question to the Blackboard Forum. This way all students can benefit from the answer, which I will do my best to reply to as soon as possible.
- ***Questions about R Programming:***
  - Can't help much when your R code/script does not work.
  - Use Google, GitHub, StackOverflow, and the built-in help functions to figure out your code.
  - Form peer groups to learn from each other and check each other's code.
- **Questions during Class:**
  - I encourage questions in class. Before, after, during practical sessions
- **Questions during Practical Sessions:**
  - Lots of time here for in depth questions on stats, code, analyses, etc.

Where are we going in this course?



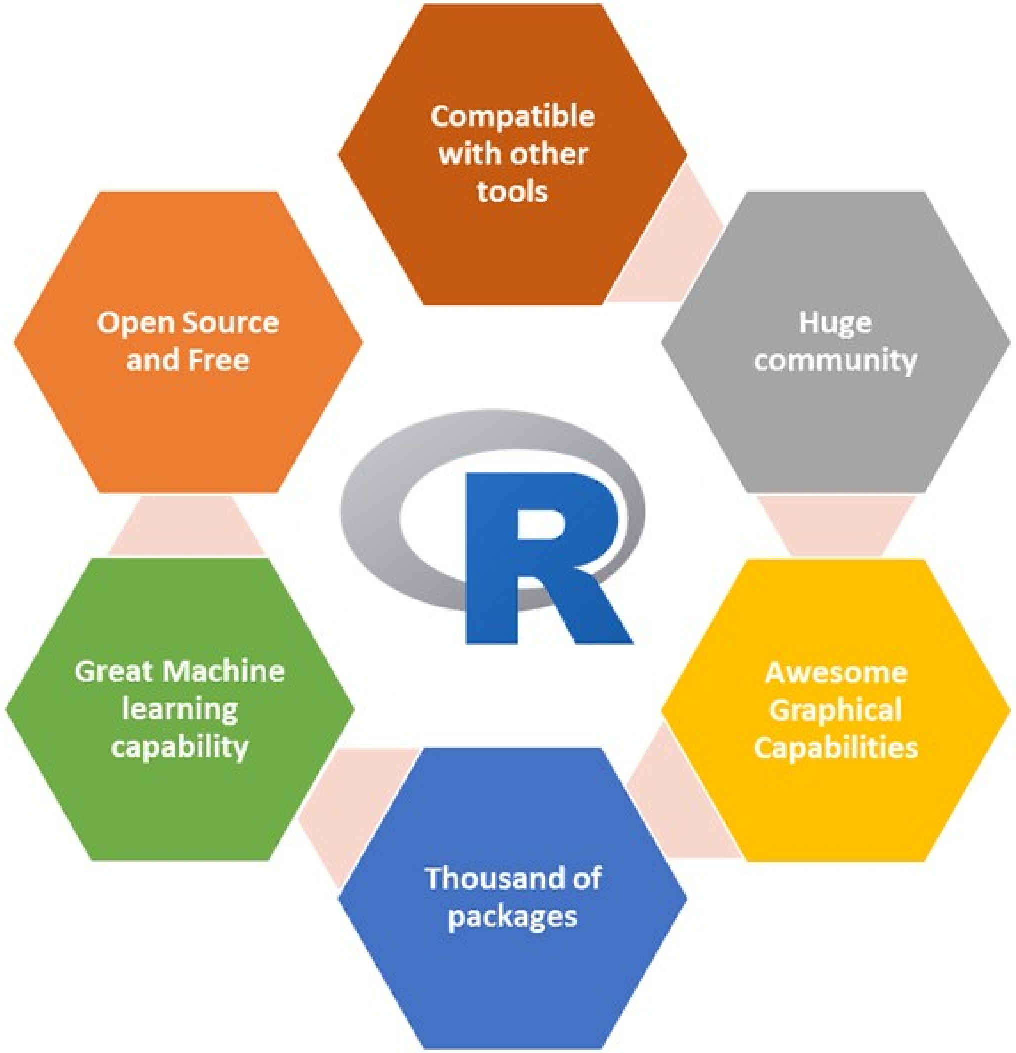
- From Andy Field

Semester Plan in Canvas

Make sure R, R studio, and Swirl working on your computers!

- Go here and follow the directions:
- https://swirlstats.com/students.html

Complete the required reading!


Compatible with other tools
Open Source and Free
Huge community
Great Machine learning capability
Awesome Graphical Capabilities
Thousand of packages

Questions?

Taco Literacy

- What if I told you a recent survey of 1000 Europeans showed that 72% of people were taco illiterate?
  - 720 out of 1000 did not understand the important anatomical components and historical lineage of tacos.
  - Would you expect 72% if you were to sample 10,000 Europeans?
  - What if we found the number went down to 43%?

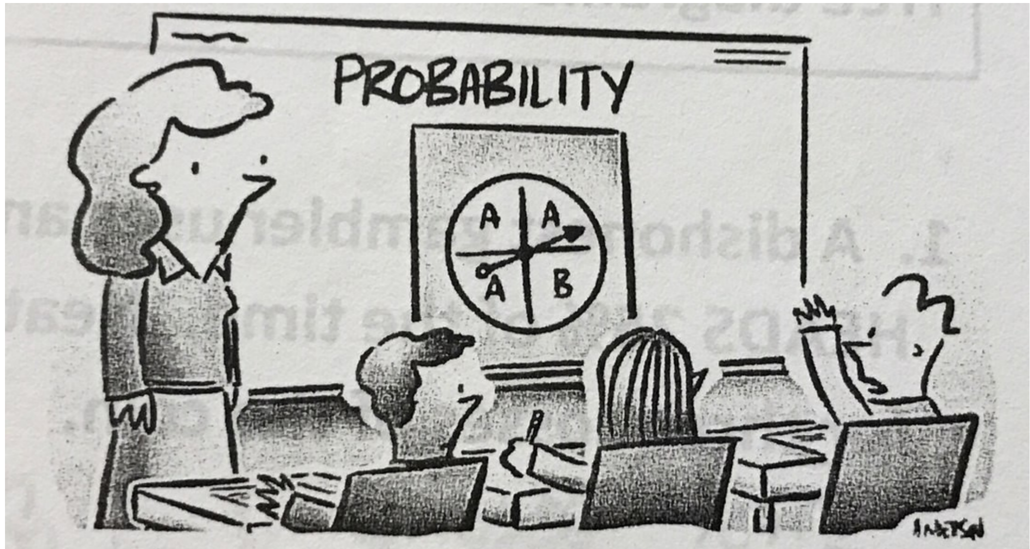
https://tacoliteracy.com/

-

TACOS
A GUIDE FROM MEXICO TO THE WORLD
A TACO IS A SOFT TORTILLA FILLED WITH MEAT AND/OR OTHER INGREDIENTS
THIS CAN BE A CORN TORTILLA
OR A FLOUR TORTILLA
THERE ARE SOME EXCEPTIONS TO THIS DEFINITION:
BURRITOS ARE NOT TACOS
QUESADILLAS ARE NOT TACOS
HARD SHELL TACOS (FOR MOST MEXICANS) ARE NOT TACOS
KINDS OF TACOS (SOME EXAMPLES)
PASTOR (PORK/BEEF OR JUST PORK)
BISTEK OR SUADERO (BEEF)
CARNITAS (PORK)
BARBACOA (USUALLY BEEF, LAMB OR GOAT)
COCHINITA (PORK)
HOW TO ORDER YOUR TACOS
5 DE PASTOR CON TODO
WITH EVERYTHING, USUALLY ONION, CORIANDER AND HOT SALSA!
NUMBER OF TACOS
KIND OF TACO (IN A TAQUERIA YOU DON'T NEED TO SAY "TACO")
EXTRA INGREDIENTS
HOW TO EAT YOUR TACOS
HEAD TILT 45°
TACO ALIGNED WITH MOUTH
HOLD FROM THE TACO'S CENTER OF GRAVITY
EXPERT MODE: HOLD THE TACOS AND SODA WITH ONE HAND... WHILE STANDING UP
HAPPY TACO DAY!
FACEBOOK.COM/PICTOLINE

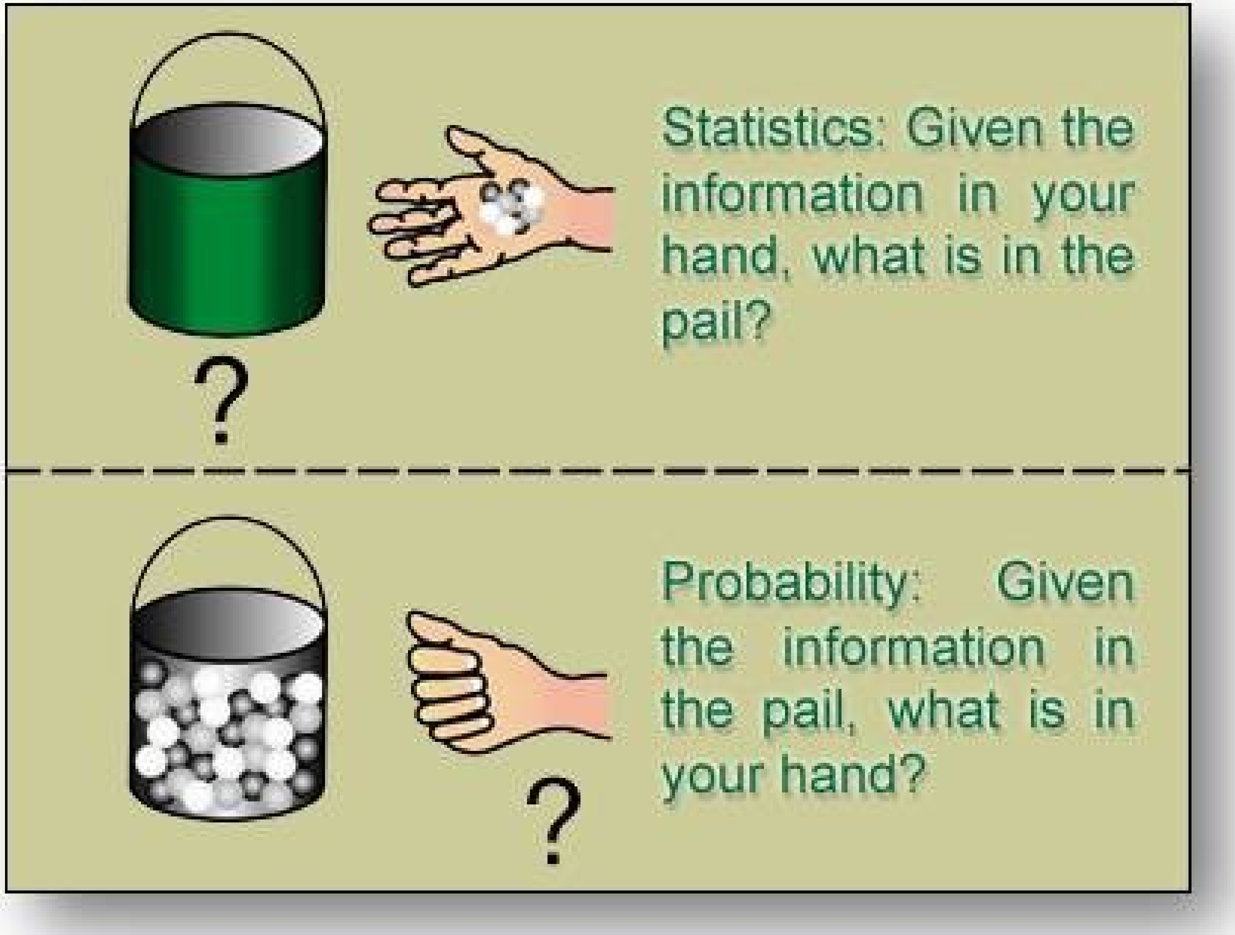
Why do we need probability theory?

- We use inferential statistics to answer questions about how representative our data are of the population
- The core of the science
- Probabilities form the basis for statistical inference.



"I know mathematically that A is more likely, but I gotta say, I feel like B wants it more."

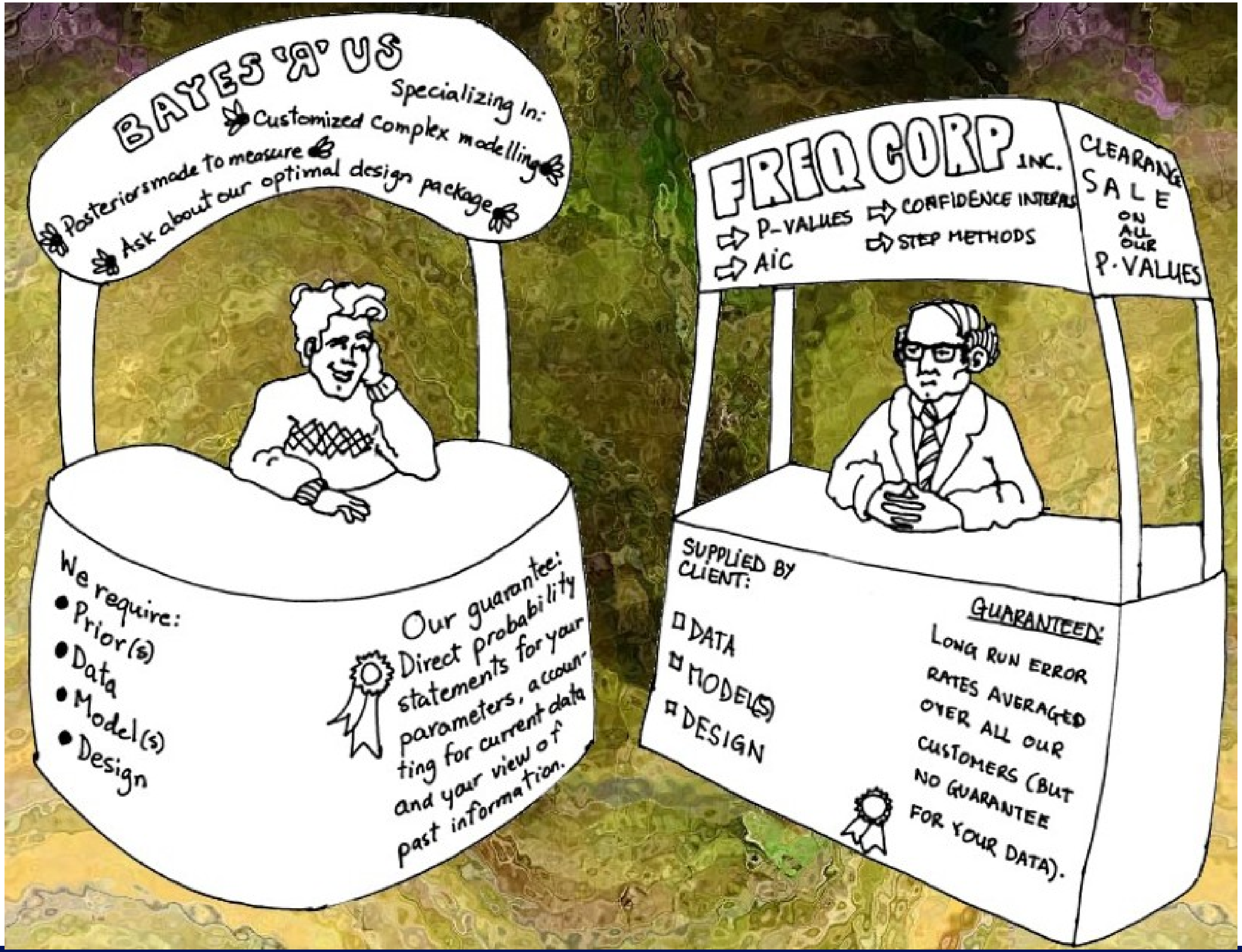
Probability vs Statistics



Statistics: Given the information in your hand, what is in the pail?

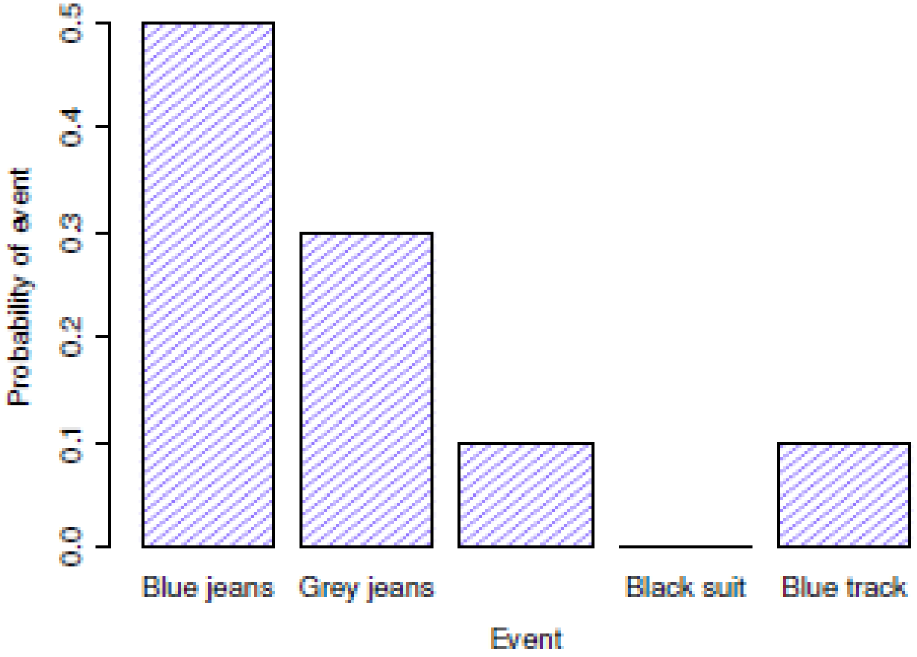
Probability: Given the information in the pail, what is in your hand?

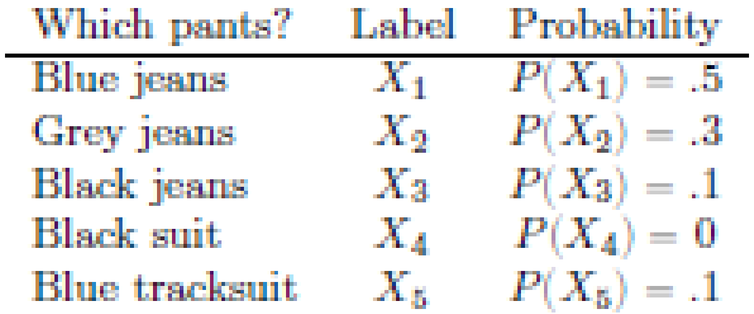
Frequentist vs Bayesian view


BAYES 'R' US
Specializing In:
Customized complex modelling
Posteriors made to measure
Ask about our optimal design package
We require:
• Prior(s)
• Data
• Model(s)
• Design
Our guarantee: Direct probability statements for your parameters, accounting for current data and your view of past information.
FREQ CORP INC.
P_VALUES
AIC
CONFIDENCE INTERVALS
STEP METHODS
CLEARANCE SALE ON ALL OUR P-VALUES
SUPPLIED BY CLIENT:
□ DATA
☒ MODEL(S)
☒ DESIGN
GUARANTEED:
LONG RUN ERROR RATES AVERAGED OVER ALL OUR CUSTOMERS (BUT NO GUARANTEE FOR YOUR DATA).

Probability Distributions

- Elementary events – For a given observation, there outcome will be one and only one of these events.
- Sample space – the set of total possible events
- P(X) = 0-1
- Sum of all probabilities is 1
- Non elementary events
  - E.g. wearing jeans
  - P(E)= P(X1)+ P(X2)+ P(X3)





Which pants?	Label	Probability
Blue jeans	X_1	$P(X_1) = .5$
Grey jeans	X_2	$P(X_2) = .3$
Black jeans	X_3	$P(X_3) = .1$
Black suit	X_4	$P(X_4) = 0$
Blue tracksuit	X_5	$P(X_5) = .1$

Binomial Distribution

- Used to model N **independent** repetitions (trials) of an experiment which has only **two** possible outcomes: Success or Failure

$$P(X \mid \theta, N)$$

- θ = success probability

-E.g., probability of tails in a coin flip, probability of six when rolling dice

- N = size parameter (number of trials)
- X is generated randomly from the distribution: $X \sim \text{Binomial}(\theta, N)$

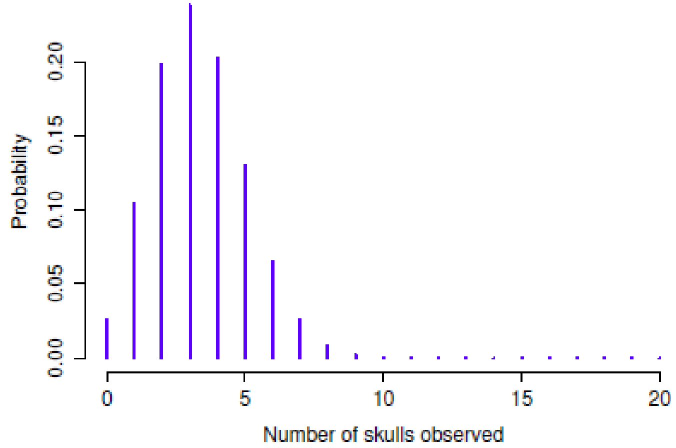
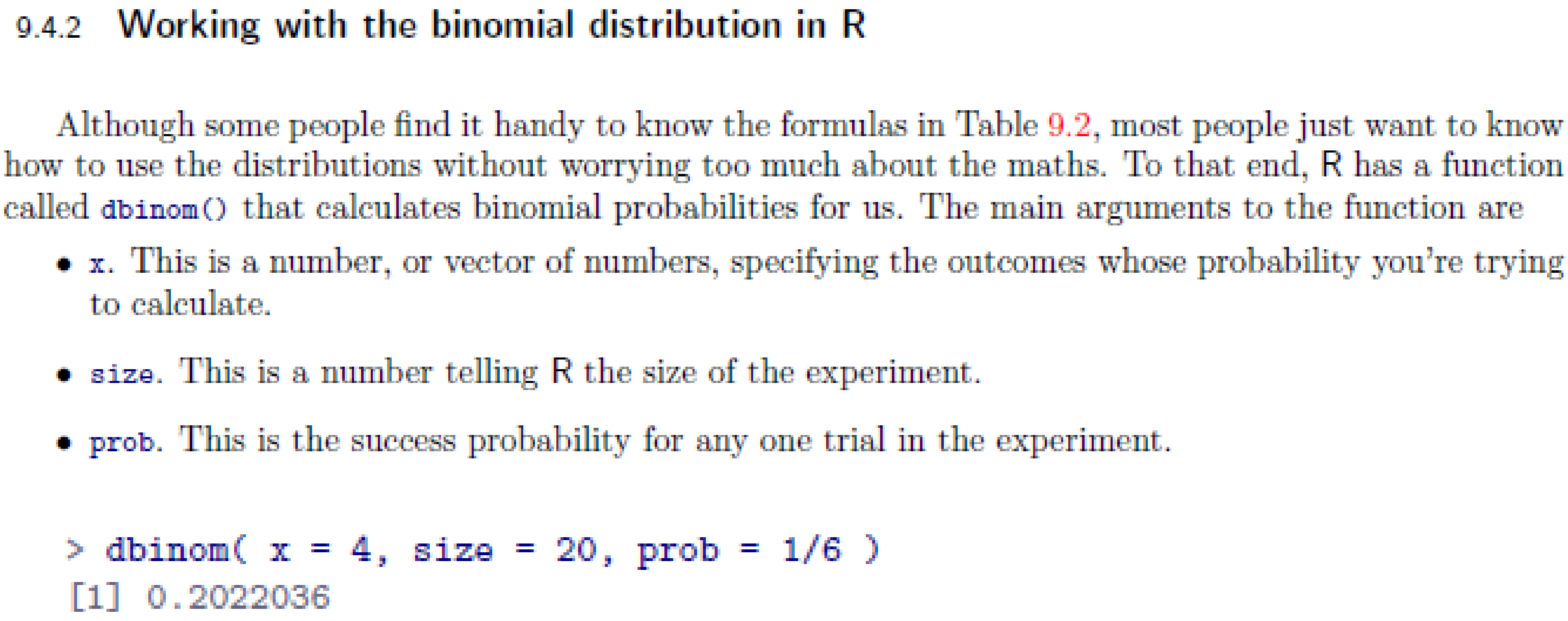


Figure 9.3: The binomial distribution with size parameter of $N = 20$ and an underlying success probability of $\theta = 1/6$. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1, and the heights of the bars must sum to 1 as well.

Binomial Distribution in R



9.4.2 Working with the binomial distribution in R

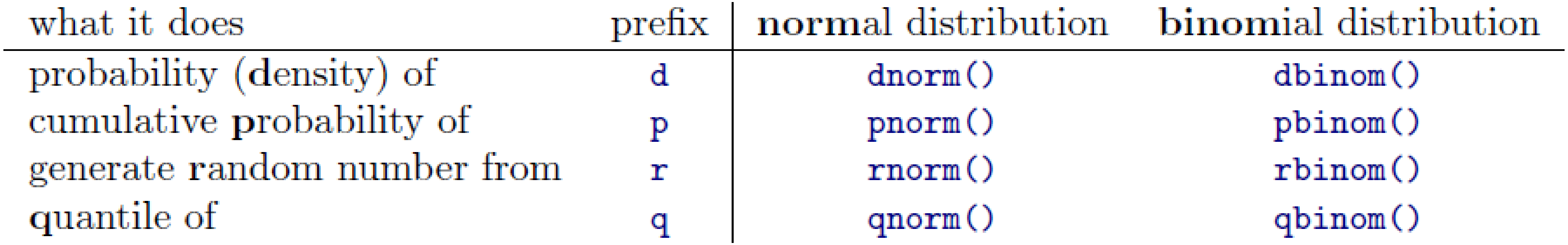
Although some people find it handy to know the formulas in Table 9.2, most people just want to know how to use the distributions without worrying too much about the maths. To that end, R has a function called `dbinom()` that calculates binomial probabilities for us. The main arguments to the function are

- `x`. This is a number, or vector of numbers, specifying the outcomes whose probability you're trying to calculate.
- `size`. This is a number telling R the size of the experiment.
- `prob`. This is the success probability for any one trial in the experiment.

```
> dbinom( x = 4, size = 20, prob = 1/6 )
[1] 0.2022036
```

- Let's try to replicate the plot from the previous slide in R!

R’s probability distribution functions



what it does	prefix	normal distribution	binomial distribution
probability (density) of	`d`	`dnorm()`	`dbinom()`
cumulative probability of	`p`	`pnorm()`	`pbinom()`
generate random number from	`r`	`rnorm()`	`rbinom()`
quantile of	`q`	`qnorm()`	`qbinom()`

- The d form calculates the probability (**density**) of obtaining a specific outcome x.
- The p form calculates the **cumulative probability**: the probability of obtaining an outcome smaller than or equal to quantile q.
- The q form calculates the **quantiles** of the distribution: the value of the variable for which there’s a probability p of obtaining an outcome lower than that value.
- The r form is a **random number generator**: it generates n random outcomes from the distribution.

Normal Distribution

- Most important in statistics!
- Called Bell Curve of Gaussian distribution
- **Continuous** distribution vs discrete case for binomial
- Described using the mean μ (mu) and standard deviation σ (sigma)

$$X \sim \text{Normal}(\mu, \sigma)$$

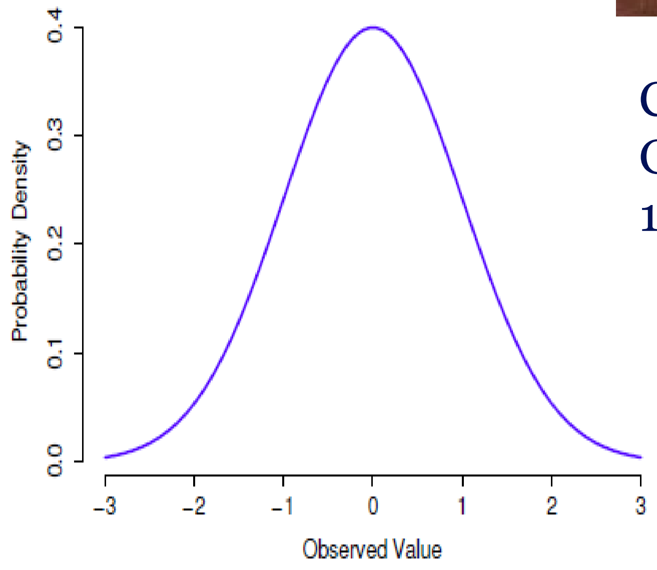


Figure 9.5: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled "Probability Density" and not "Probability". There is a subtle and somewhat frustrating characteristic of continuous distributions that makes

Carl Friedrich Gauss
1777 - 1855

Normal Distribution

$$X \sim \text{Normal}(\mu, \sigma)$$

- What happens when we change, the mean or the SD of a normal distribution?

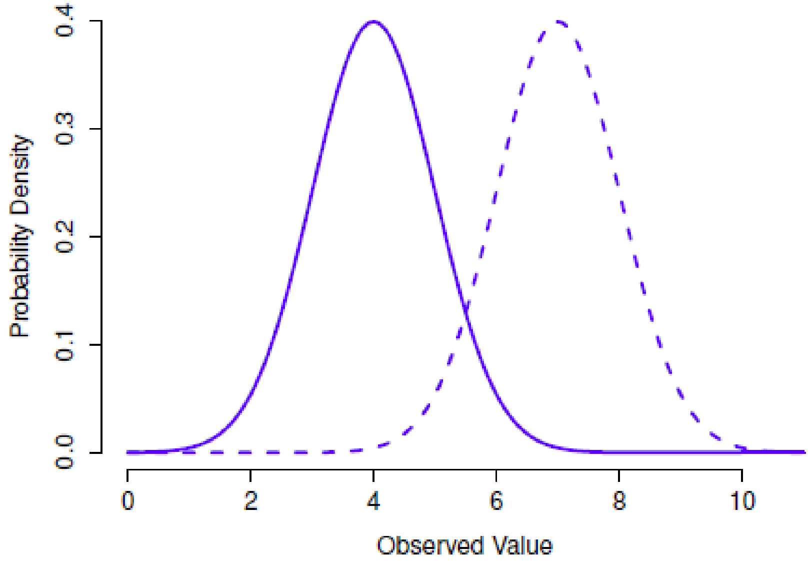


Figure 9.6: An illustration of what happens when you change the mean of a normal distribution. The solic

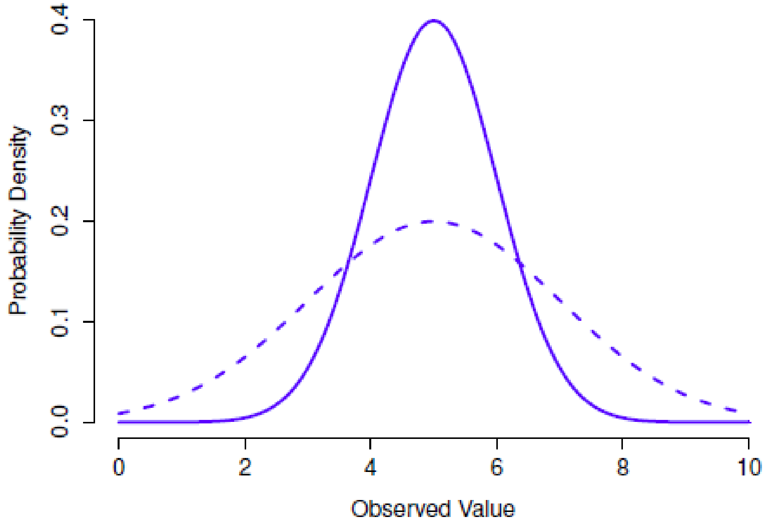


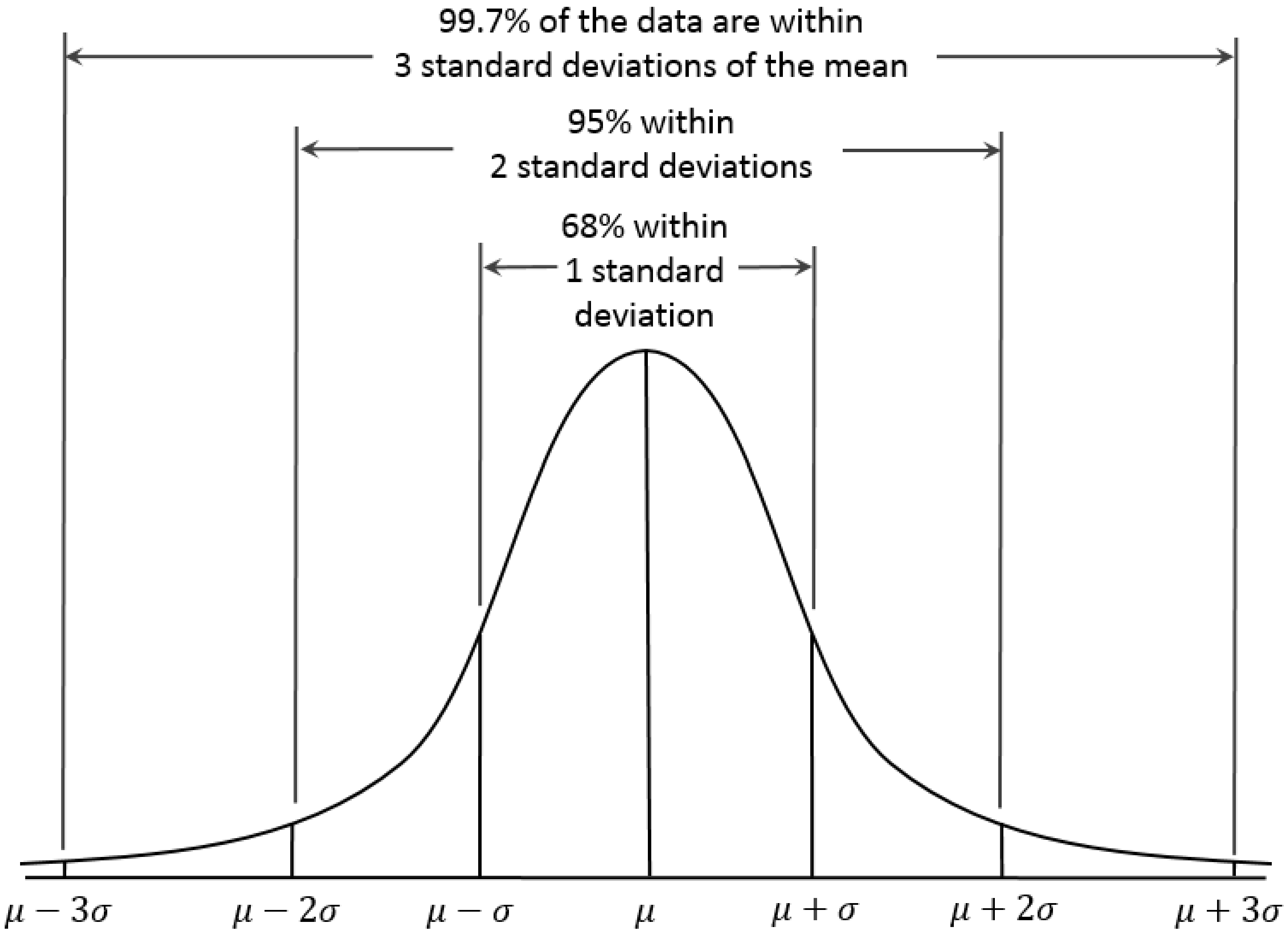
Figure 9.7: An illustration of what happens when you change the the standard deviation of a normal distribution. Both distributions plotted in this figure have a mean of $\mu = 5$, but they have different standard deviations. The solid line plots a distribution with standard deviation $\sigma = 1$, and the dashed line shows a distribution with standard deviation $\sigma = 2$. As a consequence, both distributions are "centred" on the same spot, but the dashed line is wider than the solid one.

Normal Distribution

- Y-axis describes probability density

- Probability is defined as area under the curve

- Total area under the curve must equal 1



t distribution

- Similar to normal
- Heavy tails
- Used in t-tests, regression, and more
- Used when you expect data are normally distributed, but don't know mean or SD
- R code:
  - dt(), pt(),qt(),rt()

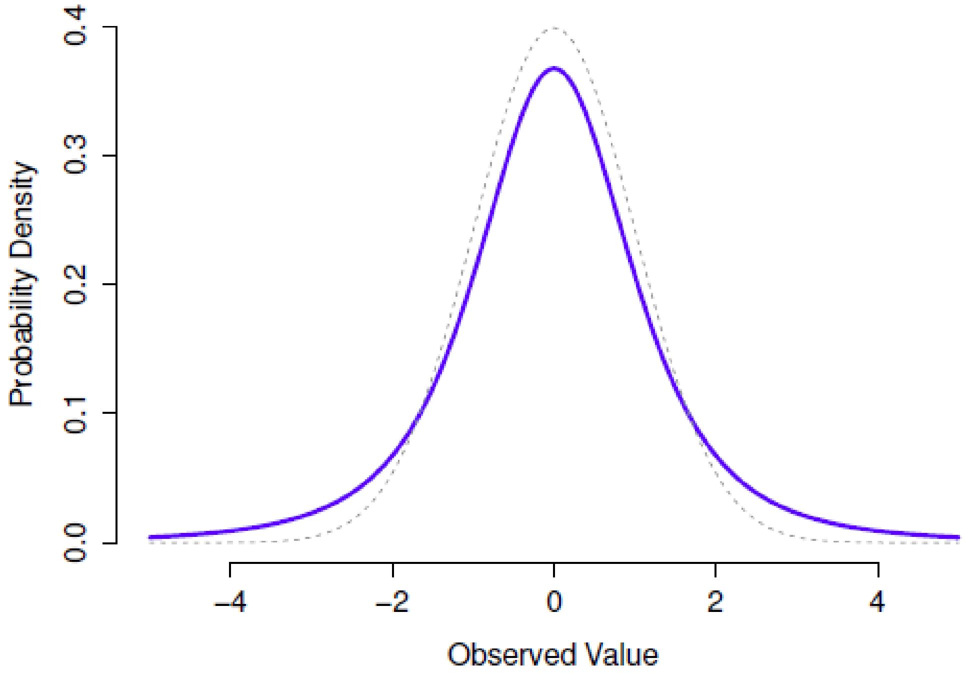


Figure 9.10: A *t* distribution with 3 degrees of freedom (solid line). It looks similar to a normal distribution, but it's not quite the same. For comparison purposes, I've plotted a standard normal distribution as the dashed line. Note that the "tails" of the *t* distribution are "heavier" (i.e., extend further outwards) than the tails of the normal distribution? That's the important difference between the two.

X^2 distribution

- Often used in categorical data analysis
- Shows up everywhere
- ‘Sum of squares’ follow this distribution
- R code:
  - dchisq(), pchisq (),qchisq (),rchisq ()

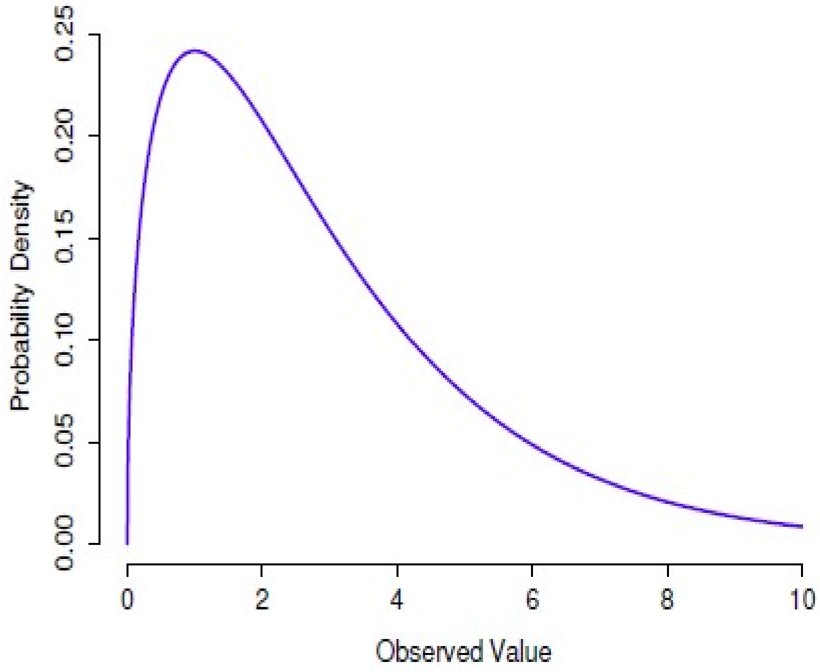


Figure 9.11: A χ^2 distribution with 3 degrees of freedom. Notice that the observed values must always be greater than zero, and that the distribution is pretty skewed. These are the key features of a chi-square distribution.

F distribution

- When you need to compare two chi-square distributions
- Comparing two sums of squares
- R code:
  - df(), pf(),qf(),rf()

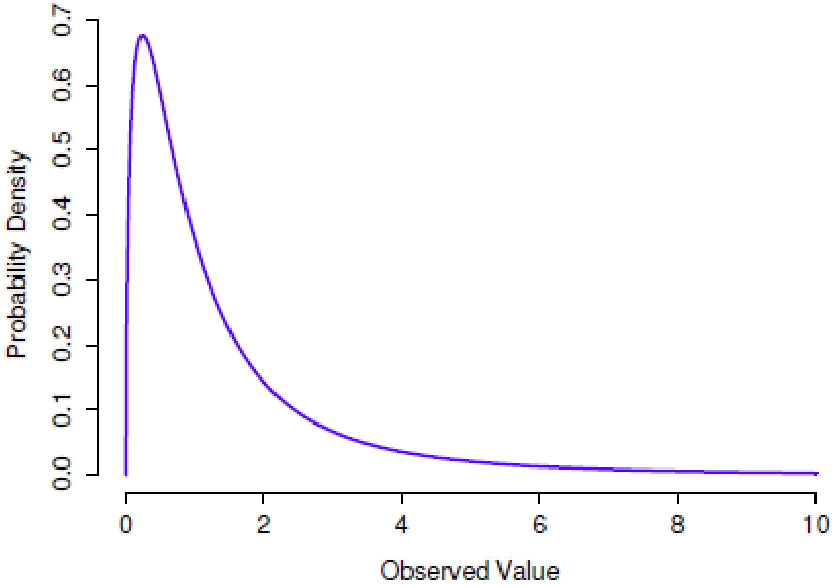


Figure 9.12: An *F* distribution with 3 and 5 degrees of freedom. Qualitatively speaking, it looks pretty similar to a chi-square distribution, but they're not quite the same in general.

Back to binomial distribution… Give it a try!

- 1) If I request 50 tacos from a restaurant that serves both hard shell and soft shell tacos, what is the probability that 25 of them are soft shell?

- 2) I want to have a pizza party and order 20 pizzas. Pizzas can either have meat on them or not meat. What is the probability that 6 of the pizzas are vegetarian friendly?

- 3) I spend too much time trying to pick a movie to watch on Netflix so I create a program to choose something for me at random. I only allow my program to choose the following genres: comedy, action, sci-fi, drama, or thriller. If I watch 50 movies using my program, what is the probability that I watch 17 comedies?

Further Reading

- Check out the following links:
  - https://www.cyclismo.org/tutorial/R/probability.html
  - http://www.r-tutor.com/elementary-statistics/probability-distributions

Thanks! See you next week! Questions?

Tilburg University