Probability and Introduction to Statistics

ALY6010

Tom Breur

Week 3, 10-NOV-2020

Agenda

- Administrative notes
- Review Discussion board
- Modeling distributions
- Choosing “the best” model
- Confidence intervals
- Sample size & Statistical power
- Preparation week 4

Administrative notes

- Several Quiz items (and assignment descriptions) are *known* to contain typos: whenever this occurs, the errors will not be substracted at the students' account
  - If you feel you have been unfairly graded due to typos, feel free to reclaim
  - In doing so, you need to provide specifics, or else I cannot investigate your claim
- These errors have been flagged and passed on to CPS faculty for future improvement(s)
  - By all means, please continue to list any errors or typos you encounter

Your TA for ALY6010 CRN 71709

Catherine Richard

Email: richard.ca@northeastern.edu

You can reach me via :

- Email
- Post on Canvas
- WhatsApp (781-526-6300)

Office Hours : Fridays 3pm

Tentative: I will send email survey & modify if certain times/days work better for students

I've worked at and am interested in :

Healthcare

Deep Learning

Start-ups

Tufts University

Northeastern University
College of Professional Studies

MPS Analytics, Statistical Modeling Concentration

I'll graduate at the end of next quarte

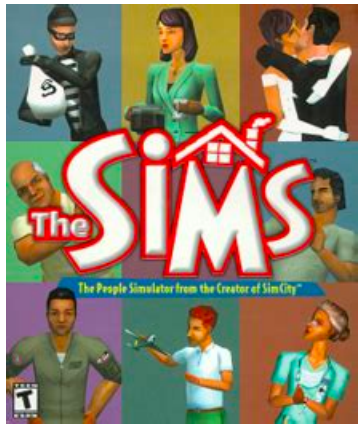
This week TA hours Thursday 9 AM & Friday 3 PM

Discussion Board

Critical review

What *is* a simulation?

- A simulation leverages a mathematical model (often in conjunction with software like R, Excel, etc.) to mimic real-world processes
- Although probabilistic calculations *are involved*, the **reference to chance processes *alone*** does not qualify to be labeled as a "simulation"
  - Simulations allow the researcher to "investigate" the behavior ofa system
  - Often used in conjunction with non-linear dynamics
- Well known examples are "management games" like MIT's "beer game", or games like the Sims, Rollercoaster tycoon, etc.



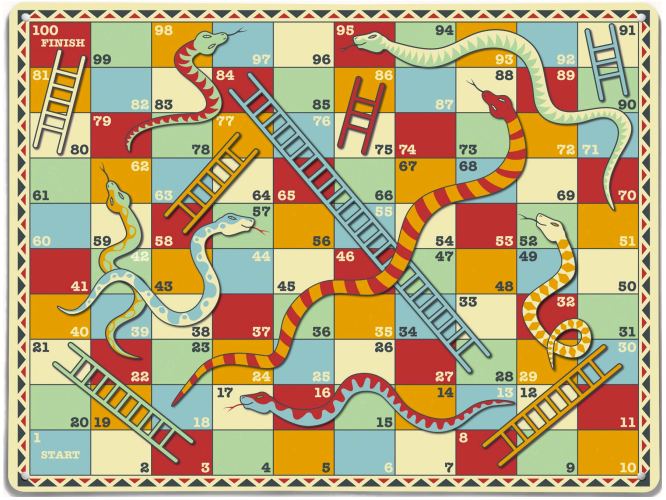
The *need* for simulations

Simulations are tyically used for one of two reasons:

- 1. Sometimes the underlying problem/process can ***not be solved*** in a so-called closed form analytical solution
  - For example: the probability distribution for determining the odds of an Electoral College win based on the odds in individual states is non-continuous; however, the state and compound odds can be approximated really well
  - Same for Cumulative Gains Curves for a predictive model
- 2. Sometimes the analytical problem ***may*** be "solvable", but we don't know this, yet (until someone provides the mathematical proof)
- Sometimes we know/think it *should* be solvable, but doing so is hard, would take too long, and not add sufficient accuracy to justify the effort (e.g. project Manhattan)

Monte Carlo simulations

- "Invented"/applied by Stanislaw Ulam & John von Neumann as a mission critical part of project "Manhattan"
- By deliberately infusing random perturbations (using random variables), the "behavior" of a system can be very closely approximated (like complex differential equations)
- Applying MC simulations allows companies to test multiple strategies (many times) when non-linearity makes it impossible to "mathematically solve" these systems
  - E.g. Snakes & Ladders
  - How many turns to finish, on average?



Modeling distributions

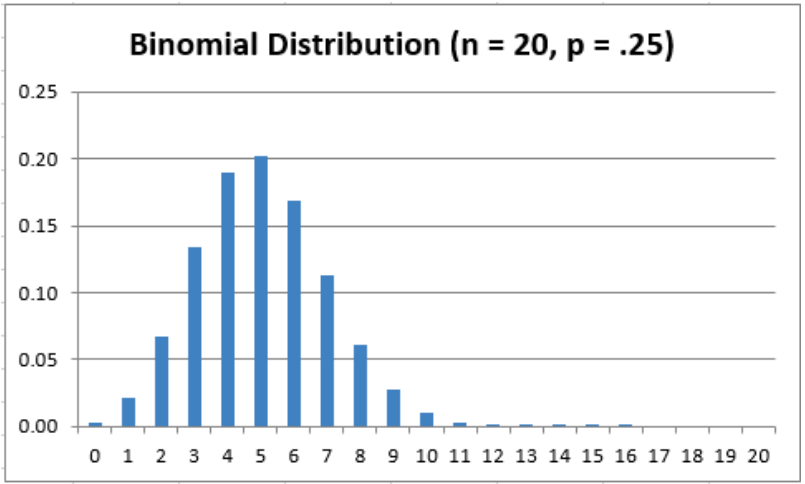
Empirical ⇔ theoretical distribution

- Assumptions about variable distributions should be considered "constraints"
- Does the underlying real-world process meet these constraints?
  - Or rather: *to what extent* does the real-world phenomenon resemble/match the theoretical distribution?
  - Note how this requires *substantive knowledge* of the real world processes that generated the data
  - Under what kind of edge conditions, might there be deviations from the desired ("ideal") distribution? Is this a pattern that can be tested in the data?

Simplifying assumptions are "normal" and OK!

Binomial distribution

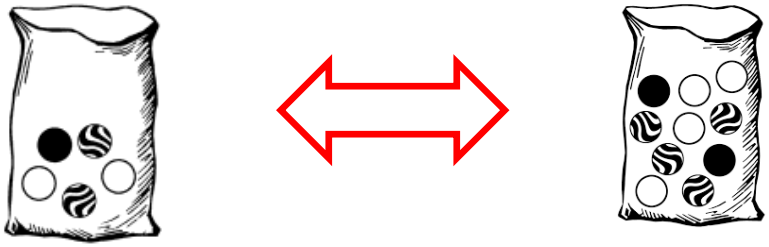
- Discrete probability distribution
- Process has *two* outcomes: “success” or “fail”
  - For this reason sometimes also called: “Bernouilli process (or trial)”
- Analogy to drawing ***with*** replacement
- Trials (“draws”) must be independent
- The odds for success/fail are assumed constant



Hypergeometric distribution

- Discrete probability distribution
- Process has *two* outcomes: “success” or “fail”
- Analogy to drawing ***without (!)*** replacement
- The odds for success/fail are assumed constant

- When the population is (very) large, relative to the number of draws, the Hypergeometric distribution approaches (“resembles”) the Binomial distribution

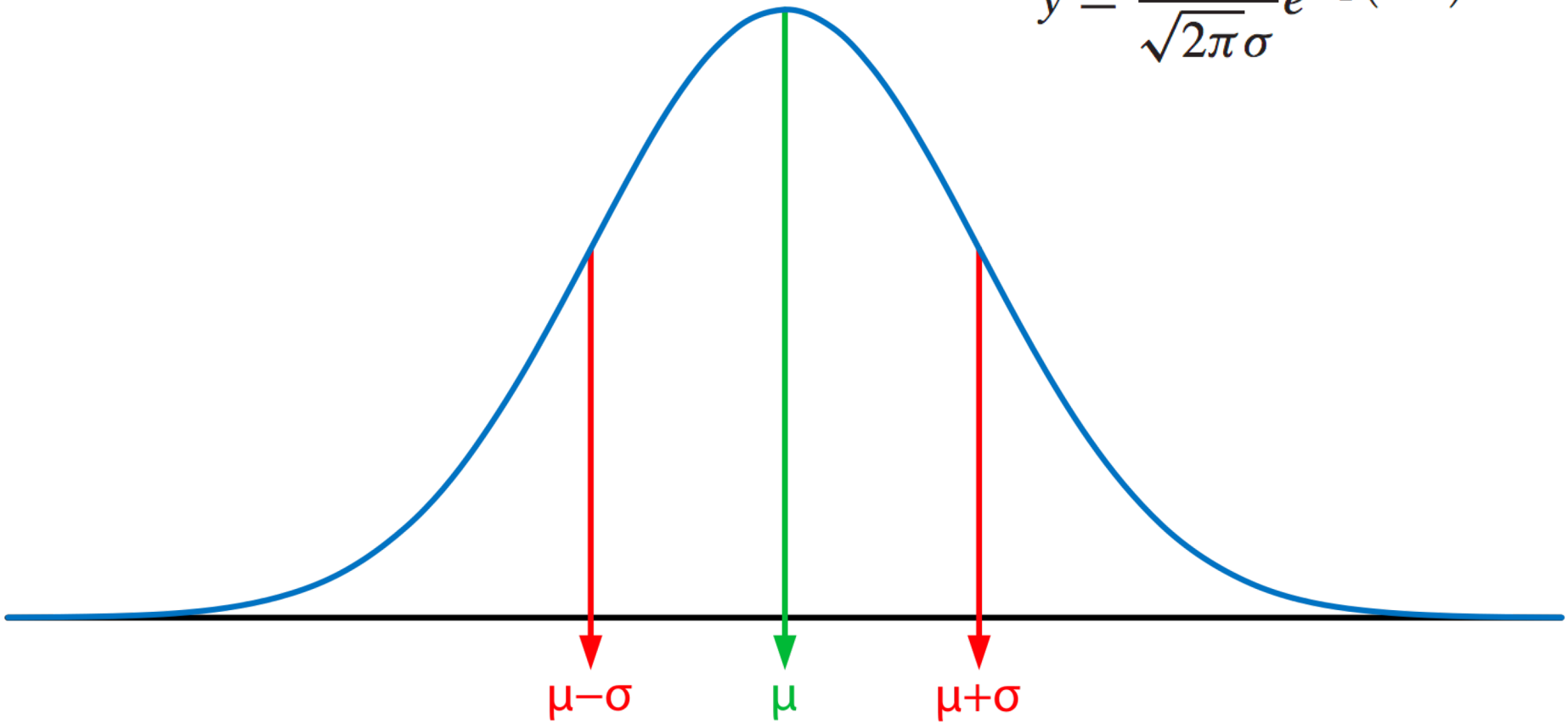


Poisson distribution

- Discrete probability distribution
- Process has a “count” (=natural frequency) as outcomes
- Typically: “count/time period”
  - Example: Geiger counter – radioactivity is count of ionized particles/second
- Phenomenon of interest *must* be independent
- Question:
  - Would this apply to a car wash?
  - To call center volume?

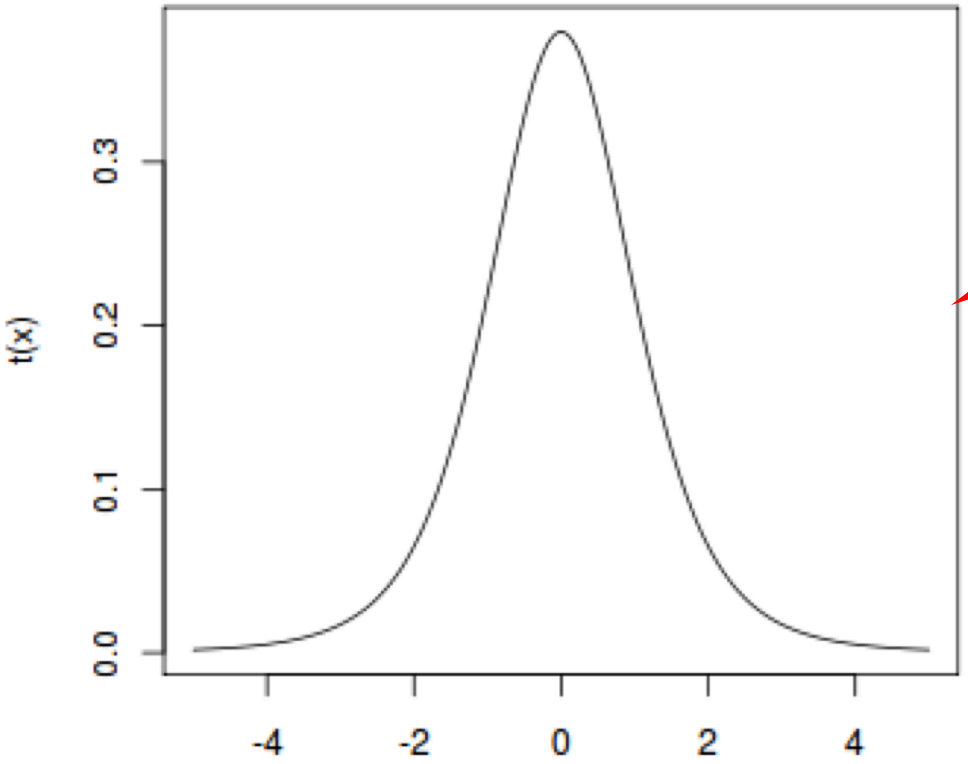
Normal ("Gaussian") distribution

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



Student's t-distribution

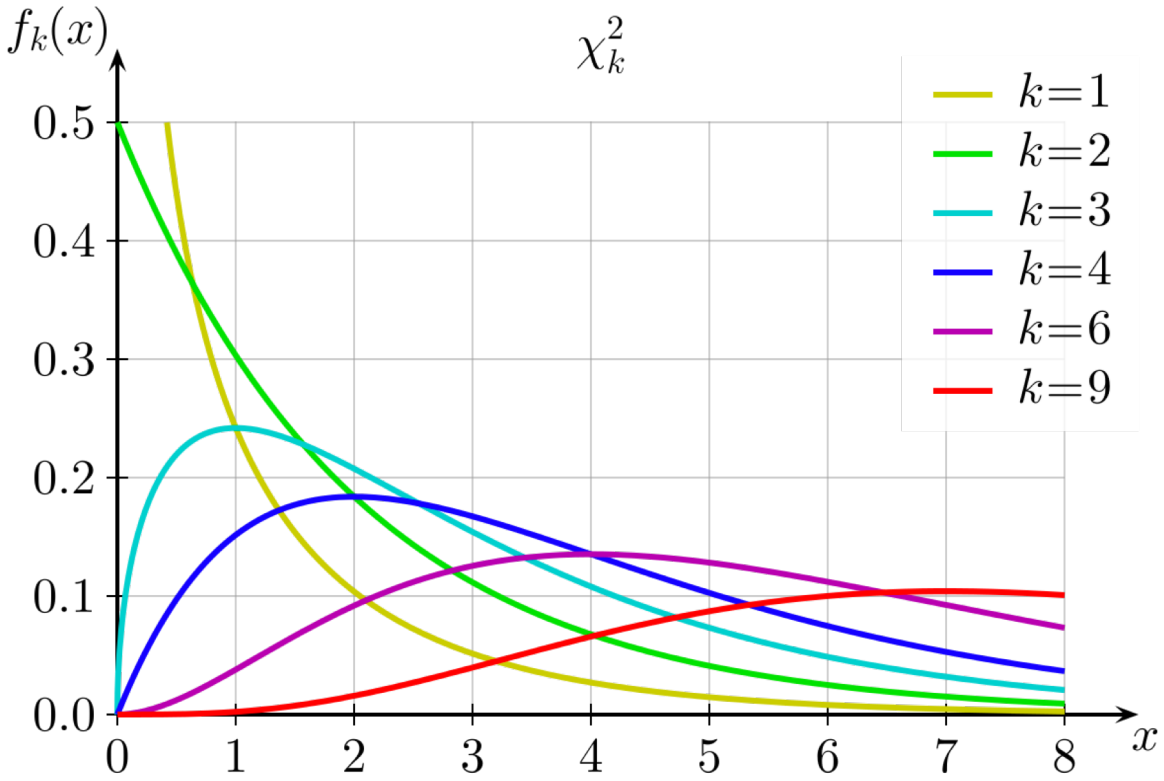
- Continuous probability distribution
- William Gosset, from the (Irish) Guinness brewery
- *Three* parameters: mean, std dev, # degrees of freedom



Visually ***almost*** indistinguishable from Z/Gaussian curves

Chi square

- Mostly used for contingency tables, but also used to test assumptions of normality
- Area = 1, for all df (the *only* parameter)



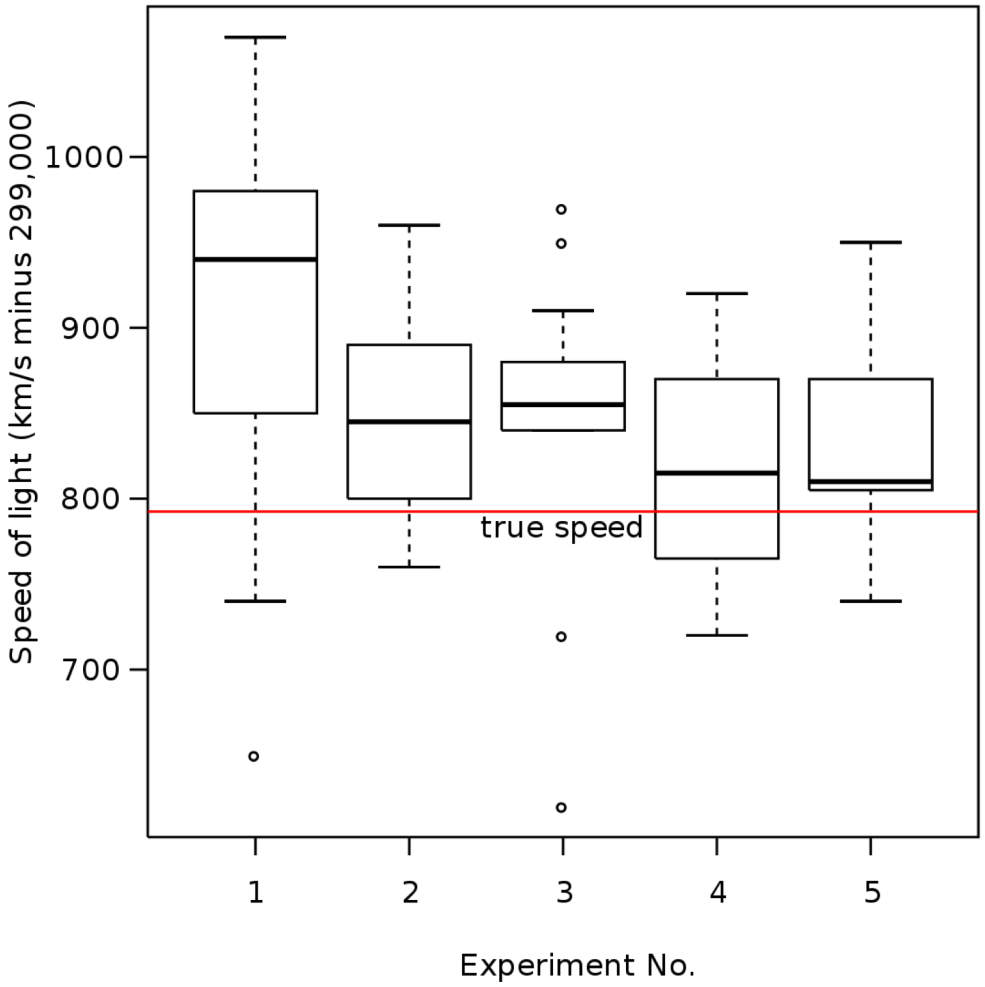
Choosing “the best” model

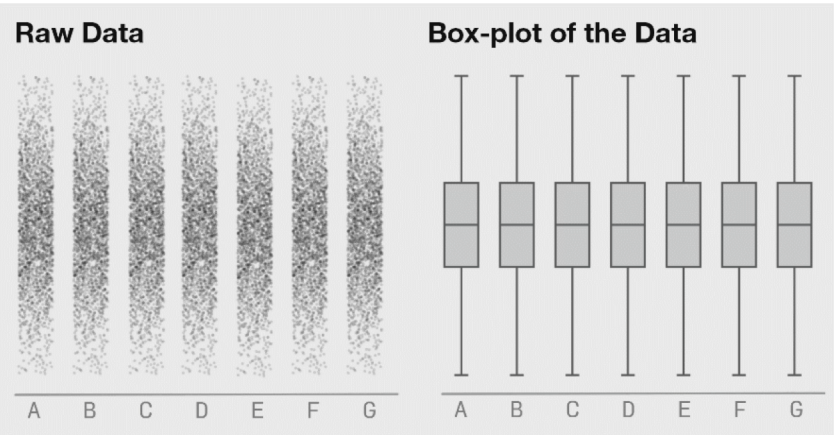
Finding "the best" model

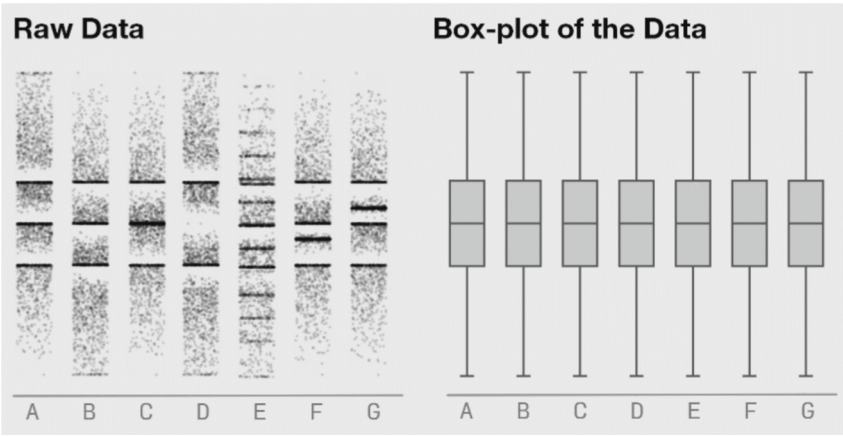
All models are wrong,
but some are useful

George Box (1919-2013)

Box plot, box & whisker diagram







Different data, **exactly** *the same* Box plot!

Some lessons to draw from this

- Often there will be multiple models that can "fit" your data reasonably well
  - There is rarely a single, "standout" winner
  - *If* there appears to be a clear winner, you have probably not framed the problem with sufficient context
- Understanding *the real-word process* that generated your data is very important
- Typical intrinsic model features that are valued:
  - Parsimony
  - Alignment with existing research and/or earlier findings
  - "Elegance" (beauty is in the eye of the beholder)

Confidence intervals

When the population (σ) is known

"Framing" the problem & tests

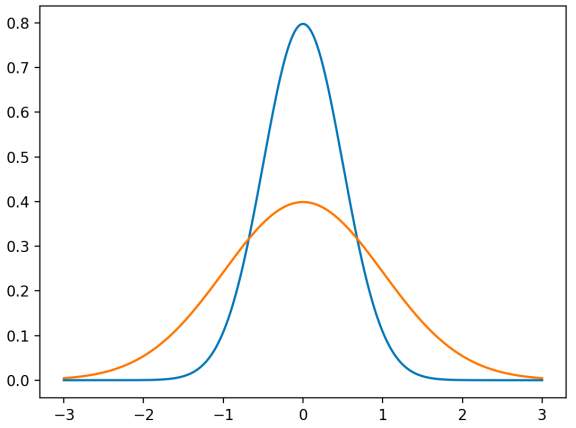
Important:

- In this week's chapter, we assume (!) that the population parameter σ is known → this determines the kind of statistical test we (can) perform
  - However, in "real world" problems, no one ever tells you what the appropriate test is, that you should use
  - As Bluman also points out repeatedly in the text, rarely will anyone mention what the core assumptions are, and/or how to test for them

Under what conditions will population parameters be "known"?

Uncertainty

- Empirical evidence in a *sample* is imperfect: every sample is subject to sampling variability
  - Sampling variability = sampling error
- Uncertainly about the ***accuracy*** of inference is reflected in the *size* of the confidence interval, and the corresponding percentages (often 95%)
- The confindence interval maps the ***range of values*** that contain the population value with, say, 95% confidence
  - Higher percentage requires a larger confidence interval, and vice versa



Assumptions

- Use of (each!) statistical test requires making certain assumptions about the underlying distributions
- Typically, there are both qualitative as well as quantitative methods to assess violations to those assumptions
- Two factors “drive” (i.e. determine) the choice of statistical tests:
  - 1. the design of the experiment
  - 2. (assumptions about the) measurement level of variables

- In this Introductory course, we limit ourselves to so-called “parametric tests”

Choice of statistical test

- Depending on the problem constraints, different tests need to be used:
  - σ is known
  - σ is *not* known (Bluman, p. 383)
- Population data may, or may not be available
- Two useful sources to determine populations data are:
  - A body of research in the field
  - Meta-analysis of prior studies (”pooling” of findings)

- This week we test against the population, soon we will be comparing *two* samples

Sample size & statistical power

The "value" of Statistical power

- Statistical power is a hugely valuable topic – notably in a business context
- The reason why practitioners care so much more about statistical power than academics is because power analysis is relevant when findings are not significant
  - Academics "only" publish significant findings, then power is immaterial
- In a business context, important questions arise like:
  - How large should our sample size (clinical trial) be?
  - What is the smallest possible effect size our study will be able to demonstrate?
  - Etc.
- https://tombreur.files.wordpress.com/2016/06/statistical-power-analysis-and-the-contemporary-crisis-in-social-sciences_201611.pdf

Iron triangle

- Three elements in (very) hyporthsis test are tied together:
  1) Sample size (N)
  2) Effect size (how large is the "real world" effect?)
  *3) Type I (& Type II) error rate – usually set at 5% in social sciences
- Between these three values, the "boundaries" of testing are confined

* Note this list is (appears) slightly different from the list in Bluman, p. 377

Preparation week #4

Requirements

- Discussion board:
  - Post contributions on *successive* (distinct!) days
  - *Minimum* of three posts, but this need not limit you
  - *First* post your primary contribution, *only then* will get access to other peoples' contributions
- Quizzes:
  - Bluman 8-1 to 8-5 quizzes
  - Module 3 R quiz
- Reading preparation week 4 (Chapter 8 Bluman)
- Reading preparation week 4 (Chapter 5-6 Kabacoff)

Discussion board: requirements

- You ***first*** (!) need to post an original contribution ("primary post") first, with a minimum (!) of 250 words
- This post needs to contain an academic reference to a reliable (!) and relevant source
  - The reference needs to be set in APA standard
- A minimum (!) of two responses are required, each 80+ words, and posted on distinct, successive dates
- All contributions need to be substantive
  - For clarity: "I agree", "I like your post", etc. do not count as substantive replies. *Instead* reason why you agree or disagree, and refer to outside soures to justify your position
- Referring to other sources or posts, or previous classes, earns "brownie points" towards top grades (100 points) for integrative learning

Discussion board: substance

- Find *two* real-life applications ("business" examples) that require/justify the use of a specific distribution
  - One continuous probability distribution
  - One discrete probability distribution
- Explain how you would determine and explain the pertinent parameters in that distribution
- How would you test whether the assumptions are tenable for that distribution?