Randomized Trials and Hypothesis Checking

Chapter 21

Performance Enhancing Drugs

Randomized trials

- The gold standard of clinical research
  - Randomized Trial or Randomized Control Trial (RCT)
- Using random sampling, participants are divided into an **intervention** or **treatment** group and a **control group**
  - **Treatment group** receives new treatment, undergoes new test, etc.
  - **Control group** receives placebo or known test
  - The control group serves as a baseline
- Samplings could be **simple** or **stratified** (chapter 17)

RCT subtypes

- Parallel groups
  - Each participant is randomly assigned
- Crossover study
  - Each participant receives the treatment over time but randomly assigned when to receive treatment
- Cluster trial
  - Groups/clusters of participants are assigned to a trial
  - Cohort study
- Factorial
  - Each participant is randomly assigned to a group that receives a specific combination of intervention and non-intervention

RCT process

- **Allocation concealment** – treatment to be allocated is not known before the participant is assigned
- **Blinded** – the participant does not know whether they are in the treatment or control group
  - May not be ethical to actually blind participants
- **Double blind** – researchers do not know which participants are in which group

RCT analysis methods

- Binary data
  - Logistic regression
- Continuous data
  - ANCOVA – analysis of covariance
- Time-to-event data
  - Censored or survival analysis

Hypothesis testing

- Create an **alternative hypothesis** and a corresponding **null hypothesis**
  - Taking *Bobiva* will make you a charming individual (alternative hypothesis)
  - Taking *Bobiva* will not make you a charming individual (null hypothesis)
- Understand the statistical assumptions about the sample
  - Both control and treatment groups will contain both naturally charming and less charming individuals
- Compute a relevant test statistic
  - Increase between pre and post test Prince Charming Assessments
- Derive the probability of the test statistic under null hypothesis
- Decide whether the probability is small enough to **reject** the null hypothesis

Rejecting the null hypothesis

- Determine a threshold, α, for rejecting the null hypothesis
  - Common values are 0.05 and 0.01
  - Could be lower and sometimes higher
- If the probability of the null hypothesis holds is < α
  - Reject the null hypothesis with confidence α
  - Accept the negation of the null hypothesis with probability of 1 – α
- Notice that we do not accept the alternative hypothesis

Testing errors

- Type I Error
  - The larger α the more likely we are to reject a true null hypothesis
- Type II Error
  - The smaller α the more likely we are to accept a false null hypothesis

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision About Null Hypothesis (H_0)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

- What has the more severe consequence?

The “Student’s” t-test

- The **t-statistic** tells us how different the estimate is from the null hypothesis
  - Measured in units of Standard Error
  - The larger the value the more likely we can reject the H_o
- T-statistic is used much like standard deviation
  - 97% of *normally distributed data* is within 3σ of μ
- We will compare to a *t-distribution* based on **degrees of freedom**

Degrees of freedom

- Describes the amount of *independent* information used to create the t-statistic
  - Generally look like normal distributions
  - Df < 30 have ‘fatter’ tails
  - Df >= 30 more like normal

Degrees of freedom example

- $variance(X) = \frac{\sum_{x \epsilon X}(x-\mu)^2}{|X|}$
- If we have 3 samples
  - We can calculate the mean
  - With the mean and any 2 values we can calculate the 3rd value
  - **So** we have 2 degrees of freedom
- For a single sample, df = # examples – 1
- For two groups, df = # examples - 2

T-tests in `SciPy.stats`

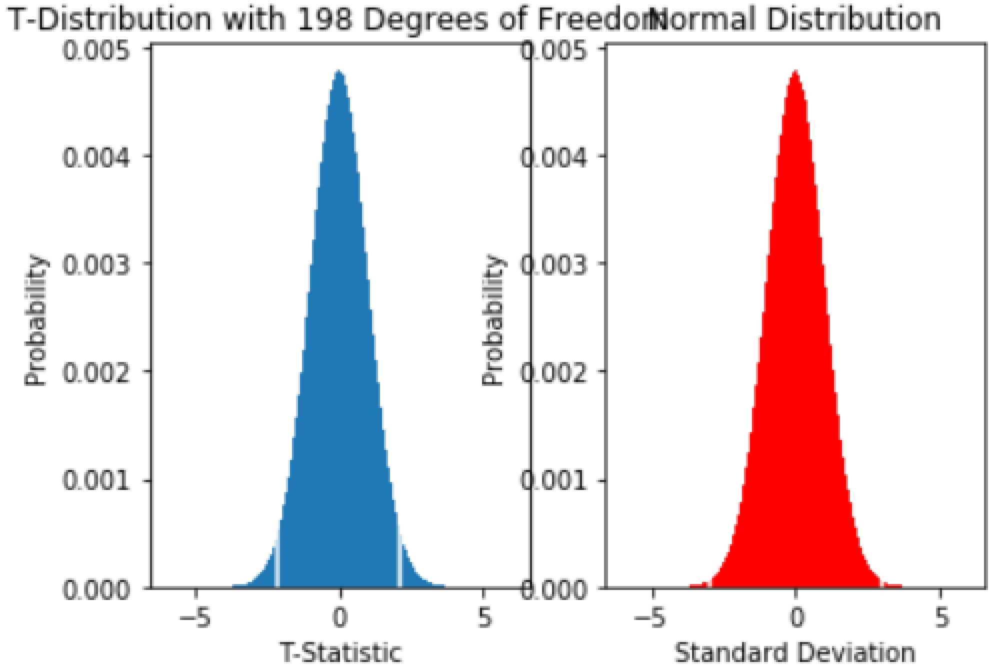
- **Independent t-test** – each observation represents a different participant
  - `stats.ttest_ind(sample1, sample2)`
- **Paired (dependent) t-test** – each participant is in each treatment (crossover study)
  - `stats.ttest_rel(sample1, sample2)`
- **One sample** – compared against a ‘known’ population mean
  - `stats.ttest_1samp(sample, pop_mean)`
- All return
  - Statistic
  - P-value

A paired test example

- We want to test a new home page
- We want 40 participants to rate the current and new pages
- Independent test
  - Recruit 80 participants
  - 40 test current, 40 test new
  - Df = 80 – 2 = 78
- Matched Pairs (dependent)
  - Recruit 40 participants
  - 20 test new then current, 20 test current then new
  - Df = 40 – 1 = 39

t vs. normal distribution

```
tstat = 2.13165598142
tDist = []
nDist = []
numBins = 1000
for i in range(10000000):
    tDist.append(scipy.random.standard_t(198))
    nDist.append(random.gauss(0, 1))
```



What can we do with the p-value?

- The p-value does **not** give the probability that the null hypothesis is True
- p applies only to the sample under evaluation
  - If p is very small then this sample has failed to support the null hypothesis
  - May have an unrepresentative sample
  - Maybe they all watched my TEDz Talk on How to be Charming!
- The larger the sample, **study power,** the stronger our result will be
- Large samples can be expensive to collect
  - Both in time and money

How many tails

- 2-tailed test
  - H_0 is there is no difference between groups
  - H_1 is there is a difference for good or ill
- 1-tailed test
  - H_0 is there is no difference between groups, or the difference in means is in a specified direction
  - H_1 is there is a difference opposite the null hypothesis
  - Weaker than 2-tailed
  - Can divide p by 2
  - Difference in means must match the hypothesis
  - Should only use if missing an effect in the untested direction is inconsequential

1-sample test

- We have a known mean for our population
- We only need to compare our sample against that mean

- Example
  - The historical GPA of SOIC grad students is 3.5
  - H_0 current applied data science students do not have a statistically significantly higher GPA than other grad students
  - H_1 current applied data science students are smarter than average
  - Get the GPA of 20 random ADS students
  - `scipy.stats.ttest_1samp(studs, 3.5)`

Skipping these – but read and understand

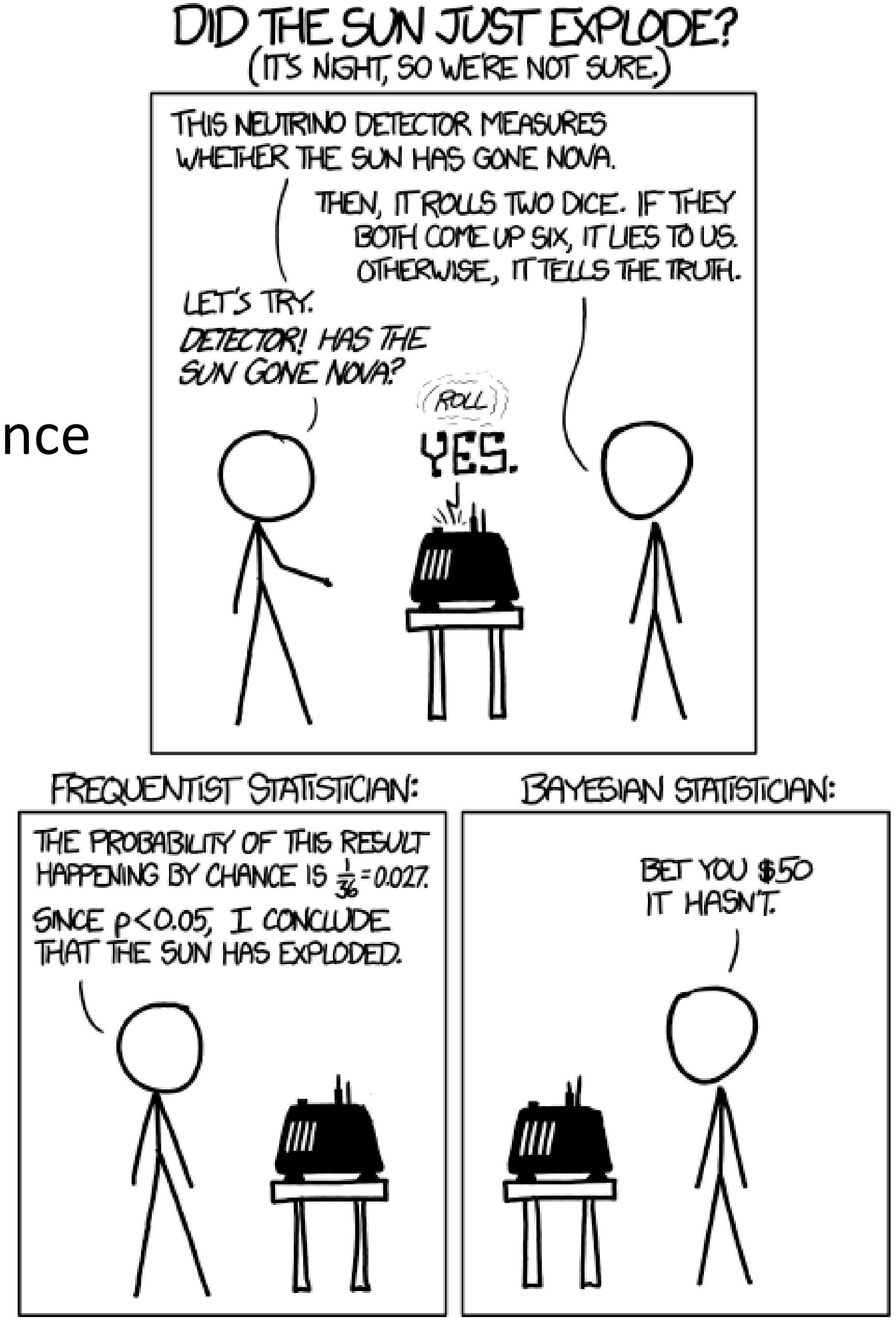
- 21.4 Words With Friends
  - Gives a good example of how running a Monte Carlo simulation aligns with the t-test
- 21.5 Come to Class
  - Size matters

Testing Multiple Hypotheses

- Are there significant differences between runners from different countries?
  - Compare female runners from Belgium, Brazil, France, Italy and Japan
    - BEL -> BRA, BEL -> FRA, BEL -> ITA, BEL -> JPN
    - BRA -> FRA, BRA -> ITA, BRA -> JPN
    - FRA -> ITA, FRA -> JPN
    - ITA –JPN
  - $\Sigma(n - 1)$ Compares
- **Bonferroni correction**
  - Applies to a ‘family’ of hypotheses
  - Divide α by m (number of compares)

Bayesian statistics

- **Frequentist**
  - Draws conclusions based on frequency of occurrence
  - Conclusions based solely on observed data
- **Bayesian statistics**
  - Utilizes additional, *a priori,* knowledge



Bayesian statistics

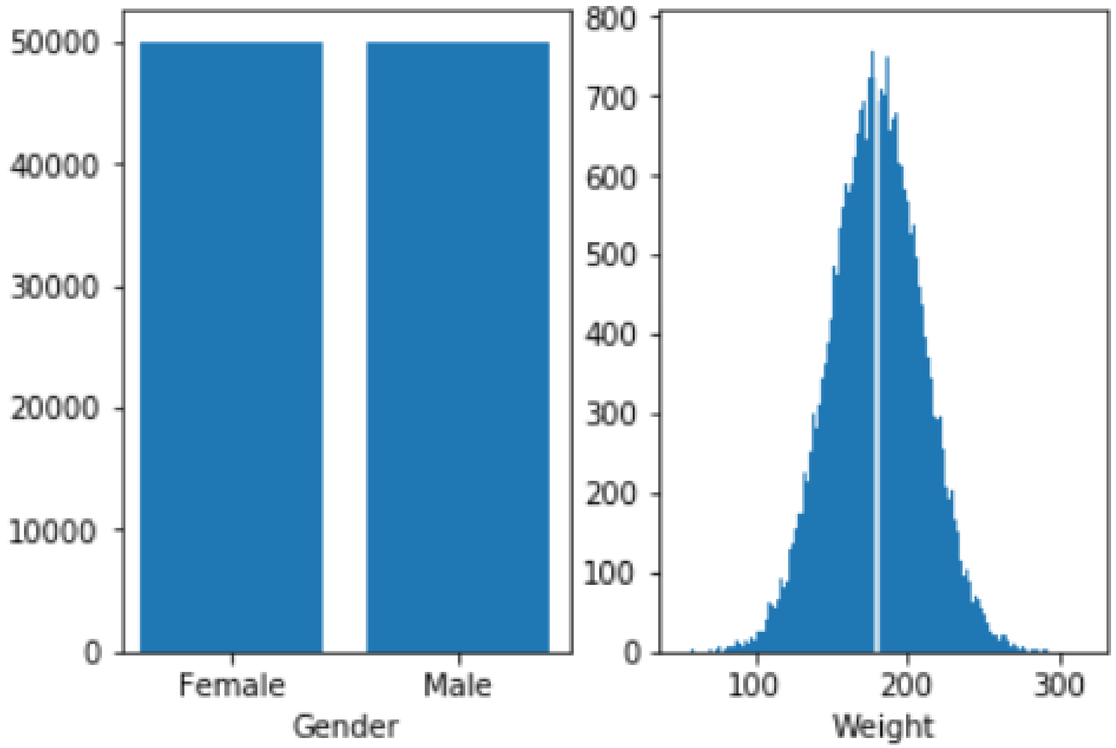
- Proposed by Rev. Thomas Bayes 1783
- Popularized by LaPlace in the 1800's

- Events are *not* independent

Example

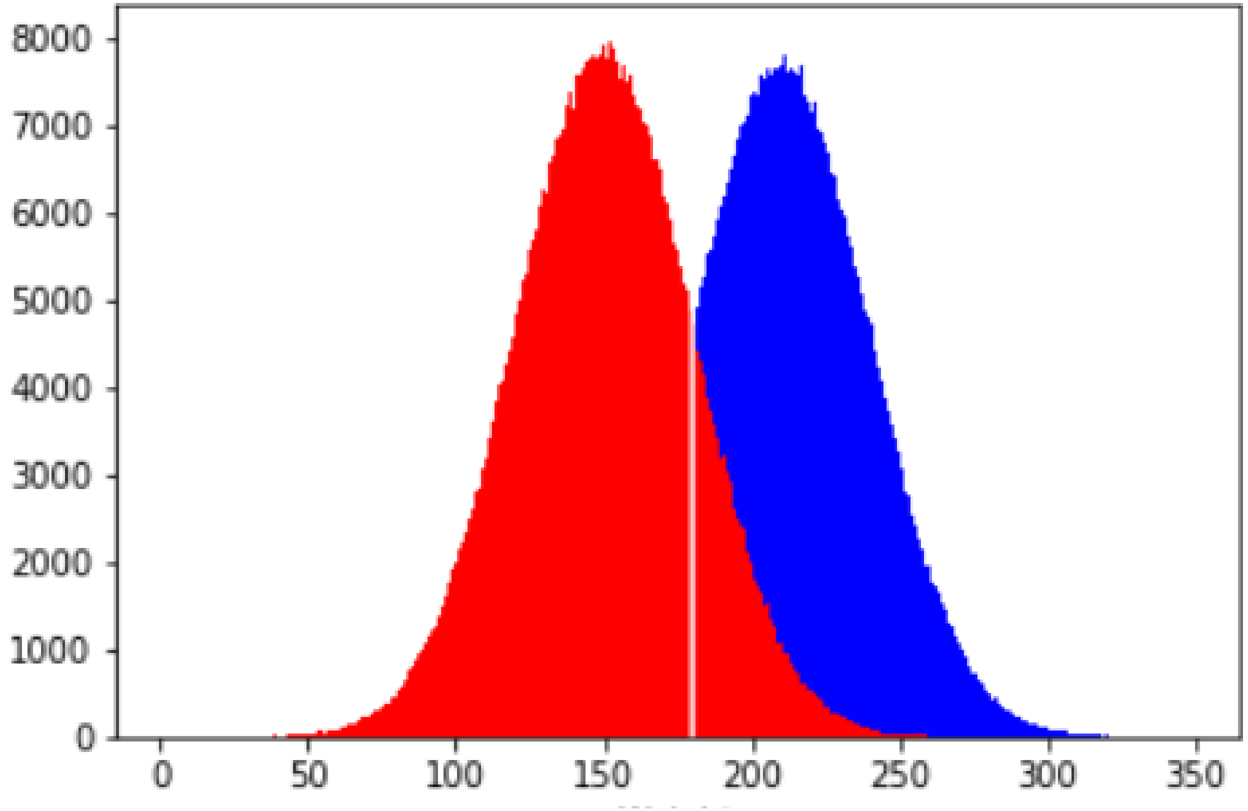
- Suppose we randomly select an adult American
- What is the probability that person is a male weighing more than 180 lb. (~82kg)?

- 0.5 * 0.5 = 0.25?



Example (cont.)

- If the selected person is male, the probability he weighs > 180 is > 0.50
- Empirical rule
  - 1 SD = ± ~ 68%
  - If mean = 210
  - And SD = 30
  - Left tail ~16%
  - So ~84% of American men weigh > 180 lb.



Conditional Probability

- P(A|B) = Probability of A given B
  - Sometimes called 'likelihood'
- $P(A|B) = \frac{P(A \; and \; B)}{P(B)}$
- P(*weight > 180*|*male*) = Probability that a male weights > 180
  - P(A) = 0.50
  - P(B) = 0.50
  - P(A and B) = 0.42 (84/200)
- So, yes, P(A|B) = 0.42 / 0.5 = 0.84

	Male	Female
> 180	84	16
< 180	16	84

- Our data was made up – most data are not symmetric
- Latest data for the US
  - 50.5% Female
  - Mean male weight 195.8
  - Mean female weight 168.5

Bayes Theorm

- $P(A|B) = \frac{P(A\ and\ B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$
- Measures “degree of belief”
- Requires “prior knowledge”

Bayesian terms

- P(A|B) = **posterior** This is the value we want to compute
- P(B|A) = **likelihood** Assuming A occurred, how likely is B
- P(A) = **prior** How likely the event is regardless of the evidence
- P(B) = **evidence** How likely the evidences is regardless of the event
  - Sometimes referred to as the **marginal probability**

Other Bayesian thoughts

- $P(\bar{A}|B)$ = Probability of *not* A given B
- $P(\bar{A}|B)$ + $P(A|B)$ = 1

April showers

- Let say there is a 30% of rain on any given day in April
- On 95% of days when it rains, dark clouds roll in in the morning
- On 25% of days with it does not rain, dark clouds roll in in the morning
- Dark clouds rolled in this morning, what is the chance that it will rain
  - $P(R|C) = ?$

April showers – what do we know?

- $P(R|C) = (P(C|R) * P(R)) / P(C)$
  - $P(R) = 0.30$
  - $P(C|R) = 0.95$
  - $P(C|\bar{R}) = 0.25$
  - What is P(C)?
    - $P(C|R)*P(R) + P(C|\bar{R}) * P(\bar{R})$
    - $(0.95*0.30) + (0.25*0.70) = 0.46$
  - $(0.95 * 0.30) / 0.46 = 0.619$
- Bring the umbrella!

Multiple conditions

- P(A|B,C) = Probability A is true given B and C are true and B and C are dependent of each other
- P(A,B | C) = Probability both A and B are true given that C is true