Week 1

Monday, December 11, 2017 8:22 PM

Suggested Reading

OpenIntro Statistics 3e, ch1: 1.1-1.5

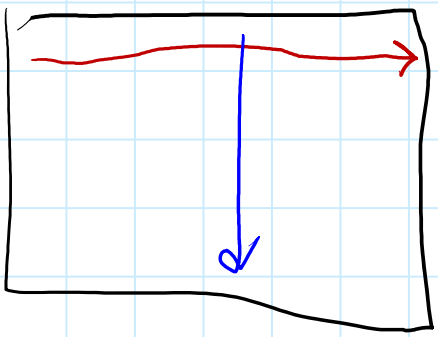
Exercises

1. 1, 3, 11, 13, 17, (9, 25, 27, 31)

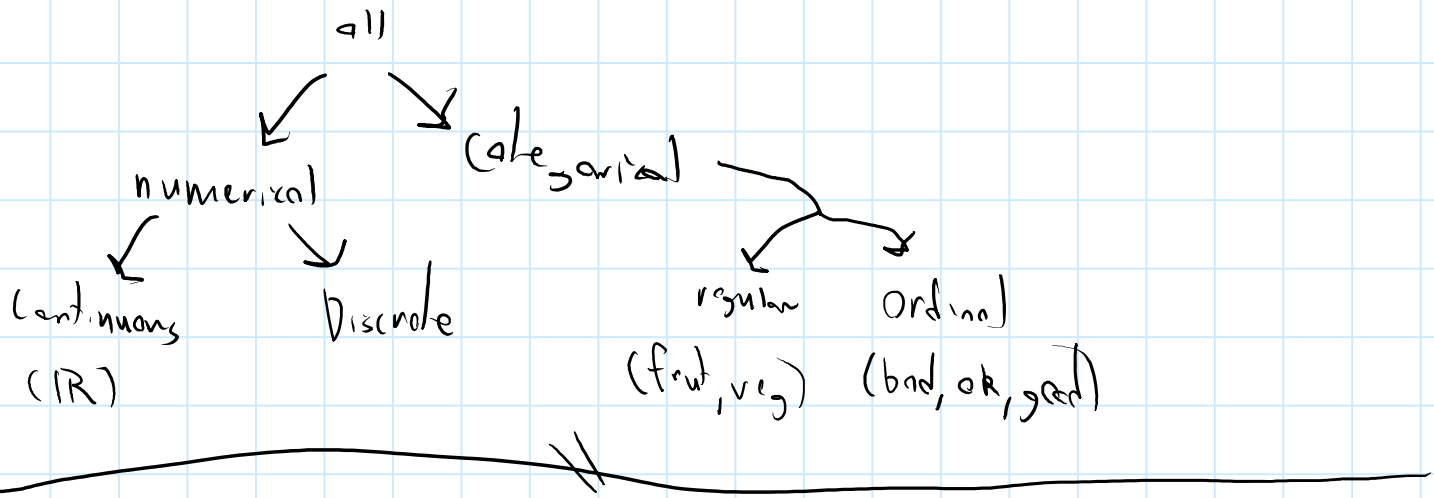
Designing Studies

Monday, December 11, 2017 9:21 PM

Data Matrix



Type of variables



Observational Studies & Experiments

Observational: No interference, can only establish correlation

Experimental: Randomly assign treatments to subjects, can establish causation

Ex: Does exercise increase energy?

Poll people who do/don't exercise on energy level

Randomly assign a group to exercise or not

Confounding Variable: extraneous variable controlling correlated values
(Why correlation does not imply causation!)

Sampling & Sources of Bias

Census: poll entire population
issues: hard to locate or measure people's data won't be recorded
populations in constant flux

ex. Cooking a pot of soup (Census is eating the pot!)

Exploratory analysis: Tasting a spoonful
Representative sample: The spoonful
Inference: deciding to add salt

ISSUE: what if spices had settled at the bottom? (Bad sample)
To prevent, stir (randomize)

Sampling Bias:
Convenience - Easily accessible individuals only ones included
Non-response: Non-random fraction of sample responds (only)

Voluntary response: Opinionated people respond more likely

Sampling Methods

Simple Random Sample - every case equally likely to be selected
Stratified Sample - Divide pop into homogeneous strata, randomly sample from these
Cluster - Divide pop into clusters, randomly select clusters, sample all w/in selected
(heterogeneous)
Multi-stage - Cluster, but randomly sample within clusters

Experimental Design

Blocking ex. Do energy gels make you run faster?
ISSUE: pro + am might be affected differently

[illegible] ex. [illegible] energy gels make [illegible]

ISSUE: pro/am might be affected differently

Sol'n stratify into pro/am

Blocking vs Explanatory

Explanatory can be imposed on an experiment (e.g. the energy gel)

Blocking intrinsic (e.g. pro/am or M/F)

Week 2

Wednesday, December 13, 2017 2:25 PM

Ch. 1 sec 6-8

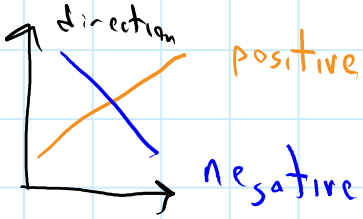
Exercises: Ch 1 39, 41, 45, 49, 51, 55, 59, 63, 65, 67

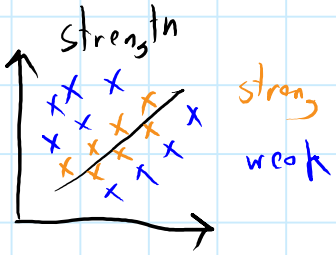
Exploring Numerical and Categorical Data

Wednesday, December 13, 2017 2:25 PM

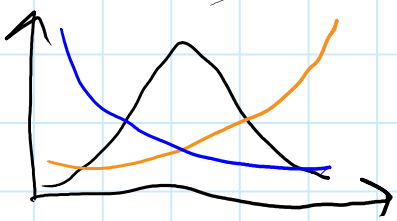
Visualizing Numerical Data

Correlations





Skewness (relative to tail)

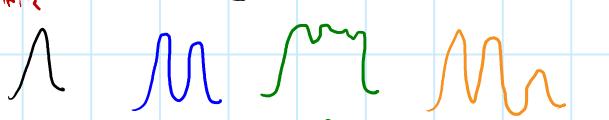


right skewed >1 (mean > median)

symmetric ≈ 1

left skewed <1

Modality



Measures of Spread

Variance: roughly avg square deviation from mean

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

why? Gets rid of negatives to avoid cancellations

Amplify large deviations

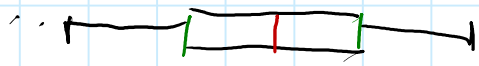
Standard deviation:

$$s = \sqrt{s^2}$$

Variability ≠ Diversity

{1,1,1,5,5} {1,2,3,4,5}

IQR = $Q_3 - Q_1$



Robust Statistics

- measures which extreme observations have little effect

	robust	non-robust
centre	median	mean
spread	IQR	SD

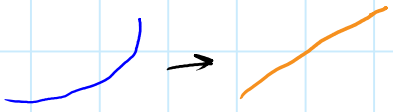
Good for skewed distributions

Good for symmetric distributions

Transformation

Used to make heavily skewed data more symmetric, more useful for modelling

e.g. log



Categorical Variables

Barplot > pie chart

Trend in dependent categorical variables?

Relative Frequency Segmented Bar Plot

or

Mosaic Plot ~aesthetic~

Intro to Inference

Wednesday, December 13, 2017 8:48 PM

Case Study on Gender Discrimination

Identical file given to promotion managers, gender randomly assigned

	promoted			
	0	1		
♂	21	3	24	88%
♀	14	10	24	58%
	35	13	48	

H_0 Difference was due to chance

H_A Discrimination

Probability of observing an outcome as extreme as what is represented in original data is the p-value

Week 3

Tuesday, December 19, 2017 8:43 PM

Reading: Ch2: 1, 2

Defining Probability

Tuesday, December 19, 2017 8:44 PM

Intro

Random process: know what could happen, but not what will

$P(A)$ = probability of event A

Frequentist interpretation of Probability:

$P(A)$ = proportion of times A would occur if the random process were observed ∞ times

Bayesian interpretation:

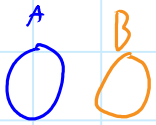
Probability is a subjective degree of belief

Allows integration of prior info

Law of large #'s:

As more observations are made, $\frac{A}{\text{all possibilities}}$ converges on $P(A)$

Disjoint Events & General Addition Rule

Disjoint ≡ mutually exclusive 

Non-disjoint = can, but don't have to, occur together 

Disjoint = mutually exclusive
Non-disjoint = can, but don't have to, occur together

For disjoint events, $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

Non-disjoint: $P(A \text{ or } B)$

General Add'n = $P(A) + P(B) - P(A \cap B)$

Sample Space:
Collection of all possible outcomes of a trial

Probability distribution:
All outcomes in sample space
AND the probabilities they will occur

rules:
Events listed must be disjoint
$0 \leq P_i \leq 1$
$\sum_i P_i = 1$

Complementary events:
2 mutually exclusive events whose probability $\Sigma = 1$

e.g. 1 coin toss
A = head, B = tail

2 coin tosses
A = HH, B = Σ of the rest (HT or TH or TT)

Independence

2 processes are independent if knowing the outcome of one provides no useful info on the other

e.g. 2 coin flips

Checking for independence:

$P(A|B) = P(A)$, A & B are independent

$P(A|B)$: probability of A given B

If data suggests dependence, do a hypothesis test!
or, use inference first:

observe diff is large, stronger evidence diff is real

if sample is large, even small differences suggest it's real
(due to law of large numbers)

If A & B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

e.g. $P(T,T) = P(T) \times P(T)$

$1/4 = 1/2 \times 1/2$

$P(!A) = 1 - \Sigma$ all permutations of A

$P(\text{at least } 1\ A) = 1 - P(!A)$

Conditional Probability

Wednesday, December 20, 2017 3:19 PM

Conditional Probability

$P(A|B)$

Bayes theorem: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

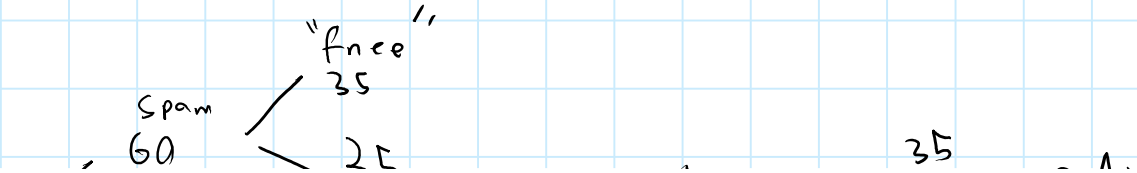
General Product rule:

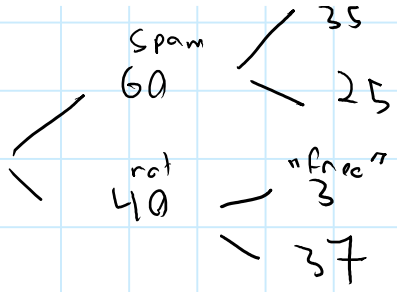
$$P(A \text{ and } B) = P(B) \times P(A|B)$$

Probability Trees

Ex. Have 100 emails, 60 are spam, 40 aren't
of the 60 spam emails, 35 contain the word "free"
of the 40 non-spam, 3 contain the word "free"

$P(\text{Spam} \mid \text{"free"})$



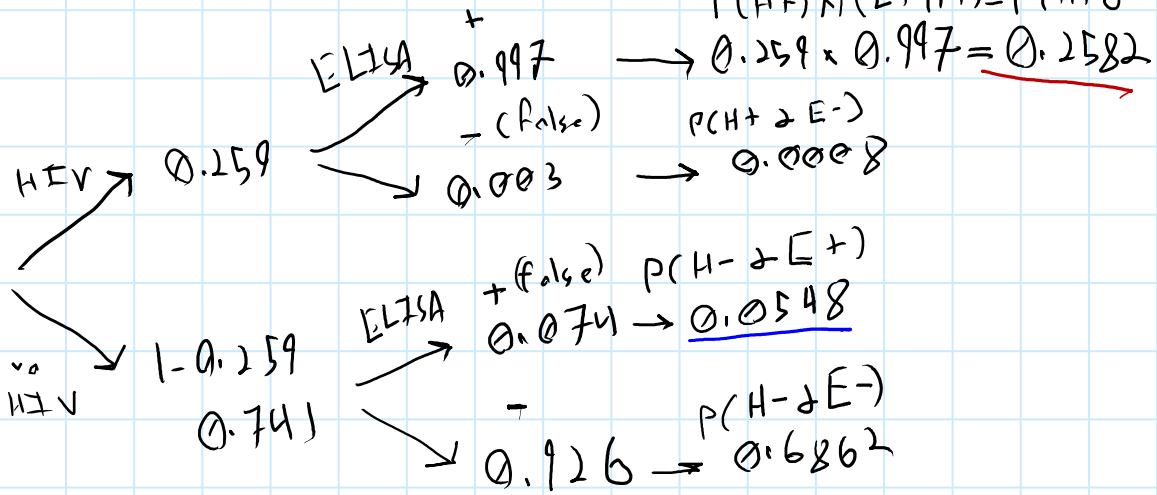


$$P(\text{spam} \mid \text{"free"}) = \frac{35}{35+5} = 0.92$$

Ex. Swaziland has 25.9% HIV infection
For those who have HIV, ELISA is 99.7% Accurate
For those who don't, it's 92.6% accurate

$P(HIV+ \mid ELISA+)$

$P(HIV) = 0.259$, $P(ELISA+ \mid HIV+) = 0.997$, $P(ELISA- \mid HIV-) = 0.926$



$$P(H+ \mid E+) = \frac{P(H+ \text{ and } E+)}{P(E+)} = \frac{0.2582}{(0.2582 + 0.0548)} = 0.825$$

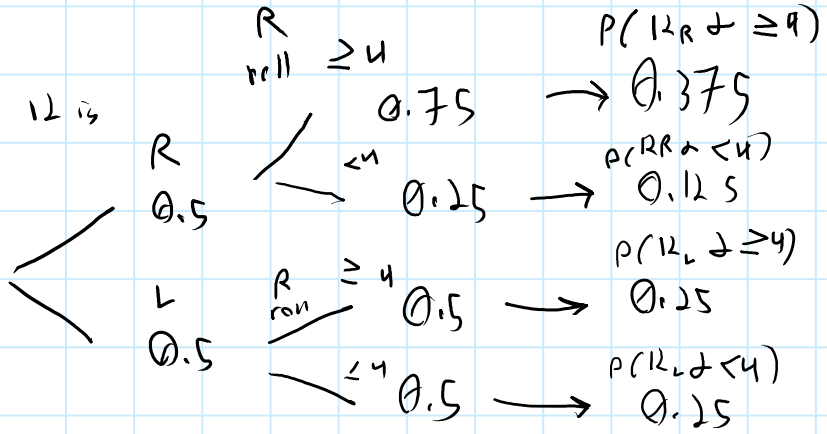
Bayesian Inference

Setup: 2 dice, 1 6 sided, 1 12 sided

Can ask to roll die 2 will be told if ≥ 4

$P(\geq 4 \mid 6s) = 1/2$, $P(\geq 4 \mid 12s) = 3/4$

Assume can't roll enough times to invoke Law of Large Numbers



$$P(12_R \mid R \geq 4) = \frac{P(12_R \,\&\, \geq 4)}{P(R \geq 4)} = \frac{0.375}{0.375 + 0.25} = 0.6$$

"Posterior probability" $P(\text{Hypothesis} \mid \text{data})$

depends on prior probability AND current observation

[Note: this is different from the probability of observed data being true ie. $P(D|H)$

Updating the prior:

The prior is updated with posterior probability from previous iteration

Bayesian method allows:

taking advantage of previous studies or physical models

- Bayesian methods allow:
  - taking advantage of previous studies or physical models
  - Integrate data as it's collected
  - Avoid using p-values, use posterior probability

Prior is important, but matters less w/ more data

Week 4

Wednesday, January 10, 2018 1:07 PM

68 - 95 - 99.7 rule

Ch. 3.1, 2, 4

Exercises

3. 3, 5, 9, 11, 17, 25, 27, 29, 33

Practice quiz

Q 6: $P(\text{identical} \mid 2 \text{ girls})$

$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B \mid A)\,P(A)}{P(B)}$

$P(2 \text{ girls} \mid \text{identical}) = 0.5$

$P(2 \text{ girls}) = 0.5 \times 0.3 + 0.25 \times 0.7$

$P(\text{identical}) = 0.3$

Quiz

Q6

$$P(\text{token} \mid \text{complain}) = \frac{P(\text{complain} \mid \text{token})\, P(\text{token})}{P(\text{complain})}$$

$P(\text{complain} \mid \text{token}) = 1 - 0.39$

$P(\text{complain}) = 0.23(1-0.39) + (1-0.23) \times (1-0.27)$

$P(\text{token}) = 0.23$

Q7 $n = 100,\ p = 0.28$

$\mu = 44.8$, $\sigma = \sqrt{np(1-p)}$

Q6 $P(\text{pregnant} \mid +)$

$$\frac{P(+ \mid p)\, P(p)}{P(+)}$$

$P(p) = 0.055$

$P(+ \mid p) = 0.99$

$P(\bar{+} \mid \bar{p}) = 0.995$

$$\frac{0.99 \times 0.055}{0.99 \times 0.055 + 0.005 \times (1-0.055)}$$

Normal Distribution

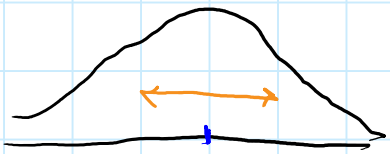
Wednesday, January 10, 2018 1:08 PM

Normal distributions

- Unimodal & symmetric

μ = avg, σ = sd

$N(\mu, \sigma)$



data encompassed

68%

95%

99.7%

Z-score: number of std deviations away from mean

$$z_x = \frac{x - \mu}{\sigma}$$

Unusual: $|Z| > 2$

Z-scores are not only used with normal distributions, but when the data is normal, a z-score can be used to calculate z-scores

In R

pnorm(val, μ, σ) = %

qnorm(%, μ, σ) = val

Evaluating the Normal Distribution

Binomial Distribution

Saturday, January 13, 2018 9:55 PM

Milgram Experiment

- measured willingness to obey evil acts (shocking someone)

2 outcomes: Don't shock (success), shock (fail)

↳ Trials w/ binary outcomes are Bernoulli variables

Binomial distribution describes P of having exactly k successes in n independent trials w/ p(success) p

$$\binom{n}{k} p^k (1-p)^{(n-k)}$$

$\binom{n}{k}$ = # scenarios = choose $\frac{n!}{k!(n-k)!}$

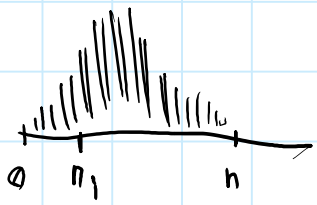
Expected value (mean) of binomial dist $\mu = np$

Std. dev $\sigma = \sqrt{np(1-p)}$

Normal Approximation

Can treat many samples which resemble normal distribution as such for

calculations



$P(R \geq n_1) = \sum_{i=n_1}^{n} P(i)$ ← summing

R:
sum(dbinom(n_1:n, size, p))

Use μ & σ to find Z-score, get probability

Normal approx may not approx binomial dist well, accounted for by adding or subtracting 0.5 to x, depending on if its above or below normal curve

When to use normal approx

$$np \geq 10$$
$$n(1-p) \geq 10$$

$Binomial(n,p) \sim Normal(n,p)$