Lecture 1

Statistics 103 fall 2017.

Instructor:
Ethan Anderes

TAs:
Michael Bissell
Maxime Pouokam

What this course is about

1) a study of random variables and the dependence btwn two random variables.
(probability theorey)

2) inferring truth based on incomplet/random data.
(statistics)

We interact with both every day.

Although some problems associated with 1) or 2) can be solved intuitively, many can be non intuitive and require formal tools to solve.

e.g. if I bet #1 on the flip of a coin, I "expect" to win about 5 times in 10 plays.

- Easy

e.g. I go to the doctor and get a strep test... its positive.

doc says:
If you have strep the test works 98% of the time.
If you don't have strep the test works 92% of the time.

This is good to know but what I actually want to know is this:
what is the chance that I have strep?

- For this one intuition fails and we need to learn formal or mathematical tools to help us solve it.

Syllabus

- Summation notation $\sum_{i=1}^{n}$
- Random variables
  - Characterizing them with PMF or Box models
  - $E(X)$, $\text{var}(X)$, $\text{sd}(X)$
- Dependence btwn random variables
  - Characterizing them with the joint PMF or Box models
  - Conditional PMFs
  - $E(X|Y=y)$, $P(X \leq 2 \mid X+Y=3)$
    $\text{var}(X|Y=y)$
  - independence
  - covariance & linear properties of $E(\cdot)$ & $\text{var}(\cdot)$
- Continuous random variables
  - Characterizing them with PDFs
  - Gaussian random variables
  - Z-scores
  - CLT
- Correlation and Bivariate regression

- MSE, var & bias of an estimator.
- Estimation & Testing
  - Comparing two populations
  - regression
  - confidence intervals.

- If you want to see more details I have old course notes up online:
  github.com/EthanAnderes/STA-103-lecture-notes
- No book required. I will post these notes online.
- No graded HWKs.
  Each week I will post practice problems for the Discussion sections and for you to practice at home.
- Your grade will be based on 3 midterm exams & the final exam.

Midterm 1: Oct. 16
Midterm 2: Nov. 6
Midterm 3: Nov. 29
} the lowest midterm score dropped.

Final: Dec 14, 10:30am - 12:30am

e.g. Hypothetical scores

Midterm 1: $\frac{10}{12}$

Midterm 2: ~~$\frac{9}{12}$~~

Midterm 3: $\frac{11}{12}$

} Top 2 gives 21 out of 24

Final : $\frac{20}{24}$

Total grade out of 100:

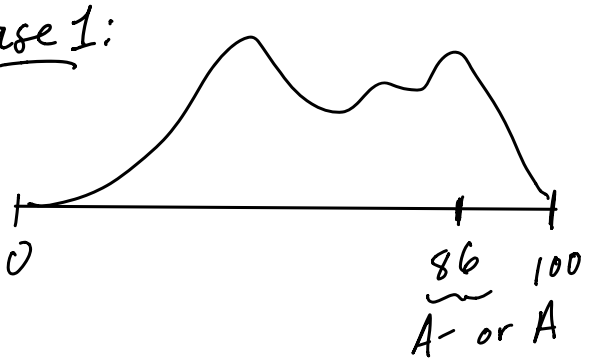
$$65\left(\frac{21}{24}\right) + 35\left(\frac{20}{24}\right) = 86$$

Midterms counts 65%

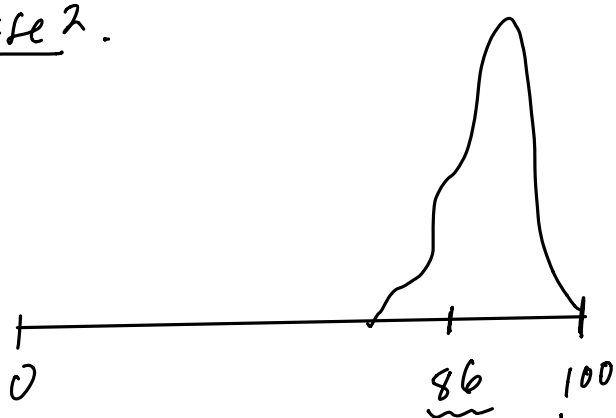
Final counts 35%

To determine your letter grade I curve based on the distribution of "Total grade"

Case 1:



Case 2.



Note: You must take the final to pass this class!!!

Summation notation

Later in the quarter we will encounter things like this:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} X_i))^2$$

summation notation

Upper case to indicate they are random numbers

We will start the class by going over this notation.

X, Y, Z, etc will generally denote R.V.s

$a, b, c, x, y \ldots$ will denote fixed non-random numbers.

$X_1, X_2, \ldots, X_n$ denote n R.V.s

$x_1, x_2, \ldots, x_n$ denotes n non-random numbers.

$\sum_{i=1}^{n} x_i$ is shorthand notation for $x_1 + x_2 + \cdots + x_n$

So $\sum_{i=1}^{n} x_i$ is just a mathy way to write $x_1 + \cdots + x_n$.

In words

$\sum_{i=1}^{n} x_i$ = "the sum of the x_i's as i ranges from 1 to n"

Note: $\sum_{k=1}^{n} x_k = \sum_{i=1}^{n} x_i$

n ← this number tells you when to stop summing

1 ← This number tells you when to start.

example

Let $x_1 = 1$

$x_2 = 5$

$x_3 = 7$

$x_4 = 2$

Then

$$\sum_{i=1}^{4} x_i = 1+5+7+2$$

$$\sum_{k=1}^{3} x_k = x_1+x_2+x_3 = 1+5+7$$

$$\sum_{k=2}^{3} x_k = x_2+x_3 = 5+7$$

example

with the same x_1, x_2, x_3, x_4 in previous example.

$$\pi \sum_{i=1}^{3} \sin(x_i) = \pi\left(\sin(x_1)+\sin(x_2)+\sin(x_4)\right)$$

$$= \pi\left(\sin(1)+\sin(5)+\sin(7)\right)$$

example

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1+\cdots+x_n\right)$$

$$= \bar{x} \leftarrow \text{the regular average of } x_1, x_2, \ldots, x_n.$$

Linear properties of Σ

We all know:

$$5(7+9) = 5\cdot 7 + 5\cdot 9$$

& $(7+2)+(9+12) = 7+2+9+12$

This is abstractly expressed with Σ as follows

$$a\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} (a x_i)$$

$$\sum_{i=1}^{n} (x_i+y_i) = \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} y_i\right)$$

Putting these together

$$\sum_{i=1}^{n} (a x_i + b y_i) = a\left(\sum_{i=1}^{n} x_i\right) + b\left(\sum_{i=1}^{n} y_i\right)$$

One application of these rules is when changing units.

example

Let $x_1, x_2, \ldots, x_{30}$ denote the daily low temp (in Farenheit) in November at the south pole.

Suppose I told you $\bar{x} = -10$

Do you have enough information to find the average temp in Celsius?

For each x_i define y_i as follows

$$y_i = \frac{5}{9}(x_i - 32)$$

y_i: the temp on day i in C°

x_i: the temp on day i in F°

We know $\bar{x} = -10$.

We want $\bar{y} = ?$

$$\bar{y} = \frac{1}{30}\left(\sum_{i=1}^{30} y_i\right)$$

$$= \frac{1}{30}\left(\sum_{i=1}^{30} \frac{5}{9}(x_i - 32)\right)$$

$$= \frac{1}{30}\frac{5}{9}\left(\sum_{i=1}^{30} (x_i - 32)\right)$$

$$= \frac{1}{30}\frac{5}{9}\left(\left(\sum_{i=1}^{30} x_i\right) + \underbrace{\sum_{i=1}^{30} (-32)}_{\text{"sum } (-32) \text{ 30 times"}}\right)$$

$$= \frac{1}{30}\frac{5}{9}\left(\left(\sum_{i=1}^{30} x_i\right) - 30 \cdot 32\right)$$

$$= \frac{5}{9}\left(\left(\frac{1}{30}\sum_{i=1}^{30} x_i\right) - \frac{1}{30} \cdot 30 \cdot 32\right)$$

$$\frac{1}{30}\sum_{i=1}^{30} x_i = \bar{x} = -10$$

$$= \frac{5}{9}(-10 - 32)$$

There are plenty of tricky problems with $\sum_i$

example

In lab (i.e. discussion.) they will show how to derive:

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

Box Models

- Box models are essentially a way to analyiz randomness without math.
- The hard part is setting up the Box model... once that is done the answer is usually easy.

e.g. I go to the doctor and get a strep test... its positive.

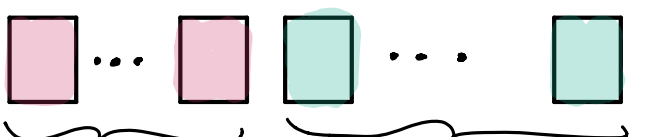
doc says:

- If you have strep the test works 98% of the time.
- If you don't have strep the test works 92% of the time.
- At any given time about 10% of the population has strep.

what is the chance that I have strep?

Consider a population of 10,000. For each person make a ticket & color it red if they have strep (green otherwise).

10,000 tickets

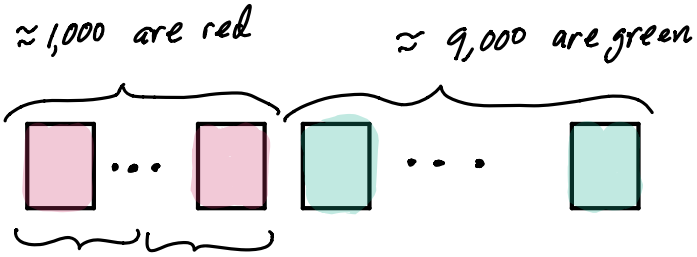


≈1,000 are red (i.e. 10%)

≈ 9,000 are green (i.e. 90%)

Now imagine giving each person the strep test.

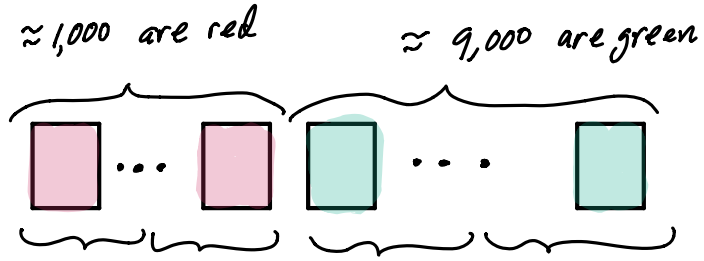
Those who have strep the test will work 98% of the time.



≈ 980 get "+"

≈ 20 get "−"

Those who don't have strep the test will work 92% of the time



≈ 980 get "+"

≈ 20 get "−"

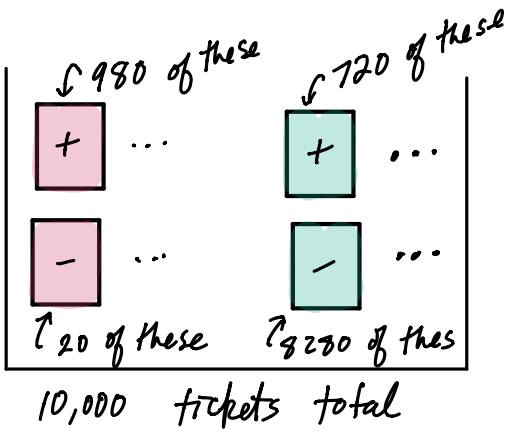
≈ 8280 get "−"

≈ 720 get "+"

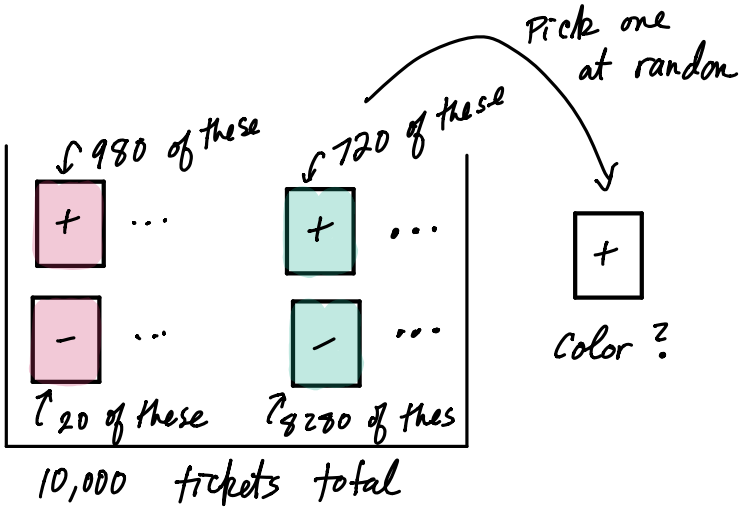
i.e. for healthy people it works 92% of the time

(8280 = 0.92 × 9000)

Write "+" or "−" on the ticket corresponding to the outcome of the strep test... then put them in a box.

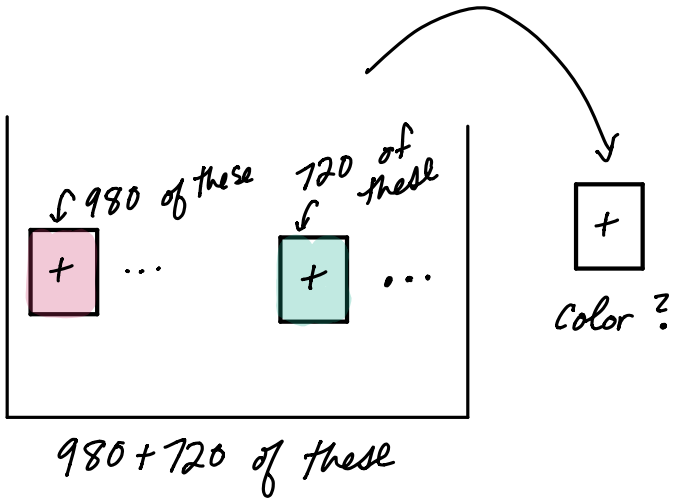


For the box model imagine someone reaches in and pickets a ticket at random. They see a "+" or "−" but not the color.



This random + represents your state of information: you got a "+" on the strep test but don't know if you have it or not.

Notice: since you got a "+", and each ticket was equally likely to pick so it behaves like a draw from

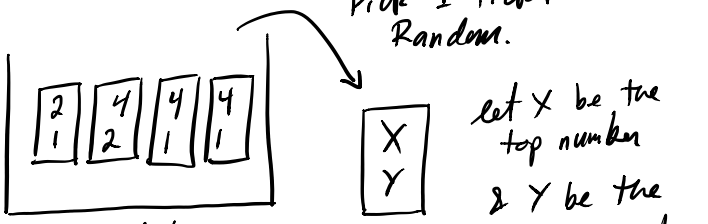


So the chance your ticket is red (i.e. you have strep) is

$$\frac{980}{980+720} \cong 0.576.$$

We will push this technique much further and use them to analyze dependence btwn random variables.

e.g.



X & Y are R.V.s

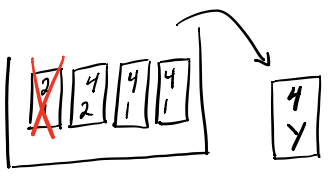
Probability calculations are very easy with Box Models.

$P(X=4)$ = "the probability that X is 4"

$= \frac{3}{4}$ ← 3 tickets with $X=4$ ← 4 tickets total.

$P(X=4 \text{ and } Y=1) = \frac{2}{4}$

$P(Y=1 \mid X=4)$ = "if someone told you that $X=4$ but not what Y is what is the chance $Y=1$".



$= \frac{2}{3}$ ← 3 tickets with $X=4$

2 of those 3 tickets have $Y=1$.