①

1) what is statistics
2) types of statistics
3) variables
4) dist of vars
5) diagram means
6) linear transformations

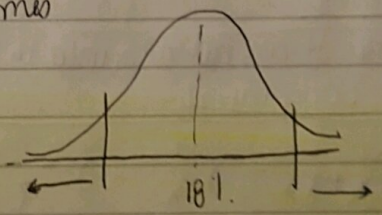
outline

SESSION 1
03/05/19

ANNOUNCEMENTS:
- sections are on saturdays - how to do things in R (in class by hand)
- double section on last week (saturday)
- mj - python ; stats - R
- attendance sheet
- project available until spring quarter schedule (optional)
- free writing counseling - U chicago grad
- flexible program → need more time etc → best
- grading is absolute
- HW à done before lecture - θ = after
- Director / Founder

WHAT IS STATISTICS?

- communicating data → descriptive statistics: summarizing data
- controlling outcomes → make good decisions in the presence of uncertainty
- uncertainty → calculated risk
  - something we don't know → variation = differences: identify → predict → control → how things depend on each other
- if no variation we don't need statistics? → not debateable -
- Event → put a # to this risk/probability (assign P(n))
  - mutually exclusive = when one occurs the other can't
  - exhaustive = cover all possible outcomes
- eg. SEO → conversion rates
  - use theoretical prob dist to assign prob in real life
- communication → bill more users
  - eg. website w terabytes of data
    - ↳ metrics
- "manage var concurrently & proactively"
  - ↳ goes after systematic variation → known & unknown causes eg. statistics → identify variation in regions
  - random variation → probability
    - ↳ inherent to the process causes



how is probability different from statistics?

① VARIATION DUE TO SAMPLING
- random = samples can be representative of pop ; no bias ; equal chance
  - > sampling variation
  - > stats : gives us # and confidence of accuracy → reduces 1 source of var (var matter)
- stratified = credit risk model (minority strata groups)
- probability : match computer w/ probable effect $Y = f(X) + \varepsilon$
  - sampling variance function (?) ← population variance
  - ε ← confidence interval

TYPES OF STATISTICS:

- descriptive statistics
  - Business intelligence
- inferential statistics → focus of program
- probability

- manage variance concurrently & proactively

variance = systematic + random

{ STATISTICS: SAMPLE → POPULATION
PROBABILITY: POPULATION → SAMPLE

- assume a distribution then explain a probability
- or observe the sample to form guess the distribution

VARIABLES:

- independent = measurements we use to explain y dependent var
- modeling frameworks
  - supervised → have an ~~#~~dependent var
  - unsupervised → no dependent var (y)

} machine learning!

discrete {
- nominal - labels only; no order; no degree → P(X)
- ordinal - w order → don't need to be evenly spaced
  - ⊛ % - bound scale → not technically ordinal?

♦ garbage in, garbage out

continuum {
- interval - don't have 0 (:) - linearly related (Age) - more analytic friendly
- Ratio - Real 0 points → Return = Ratio
  - Acting to the power of 2
  - Probability of intervals

DISTRIBUTION OF VARIABLES:

- Normal distribution (continuous) ⊛ → Read about this later
  - μ, σ
    - Arithmetic mean
  - σ = how far from the mean → "2σ"
  - σ^2 = variance → no "are" interpretation like σ
  - symmetry
    - ♦ kurtosis → Finance
    - ♦ skew

- CENTRAL TENDENCY MEASURES:
  - stem & leaf
  - mean
  - median
  - mode

} ⊛ this graph is good

median = inconvenient mathematically
- doesn't rep pop if skewed
mean = biased to outliers
mode = more ordinal

⊗ Healthins ✓ ⊗ start HW 1B
① STATS REV
② Rewsrch } ⊗
③ Prog For Anlyt }

DISPERSION MEASURES

- σ^2 ⟶ "Statisticians bread & butter"
- σ

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

⟶ square it to get a meaningful metric

"sums of square"

⟶ { L1 norm - absolute val
L2 norm - squared }

⊗ Residual sum of squares

(MSE)

⊗ Average sum of squares

- skewness
  - long tails
  - happiness is relative
  - kurtosis = extreme events sensitivity
    - platykurtic = chicago
    - leptokurtic = manila
  - modality
    - implies diff segments in population ⟶ separate to get unimodal
    - separation by age

nominal {
- Pie chart: labels must be mutually exclusive (ME), small sample size, people who like other stuff (CE)
- Frequency table
  - M = weighted average
}
- histogram

LINEAR TRANSFORMATION OF VARIABLES

$y = a + bx$

- correlation
- covariance

- trimmed mean = if data biased

03/06/19 session 1 notes/ concept Review

Ch 7: NORMAL DISTRIBUTIONS → continuous distributions

1) INTRODUCTION TO NORMAL DISTRIBUTIONS

- "bell curve", "Gaussian curve"
- μ, σ

$$\text{density of normal dist} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

e = base of ln or log(e)

Hypothesis testing:
H_0: $\mu = 100$
H_1: $\mu \neq 100$

Decision	H_0	H_1
H_0	confidence	type II β
H_1	type I α	power

- Features of normal distributions:
  1) symmetric around their μ
  2) mean = median = mode
  3) Area under normal curve = 1.0
  4) denser in center & less dense in tails
  5) Defined by 2 parameters: μ & σ
  6) 68% of its area is within 1σ of μ
  7) ~95% of its area is within 2σ of μ

2) HISTORY OF THE NORMAL DISTRIBUTION

- Abraham de Moivre - "as the # of events increased, the shape of the Binomial dist approached very smooth curve"
- many natural phenomena are naturally distributed
- Laplace = central limit theorem

3) AREAS UNDER NORMAL DISTRIBUTIONS

⊗ online normal & inverse normal calculator → wtf.
⊗ varieties demonstration

4) STANDARD NORMAL DISTRIBUTION

- ~~normal distribution with~~ a $\mu = 0$ and $\sigma = 1$ is standard normal dist.
- z table

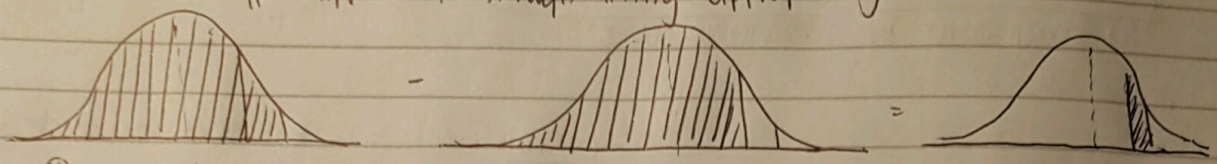
$$z = \frac{(X - \mu)}{\sigma}$$

z score
z = normal dist
X = value on original dist
μ = mean of orig dist
σ = σ of orig dist

- "standardizing the distribution" → transforming dist to std normal

5) NORMAL APPROXIMATION TO THE BINOMIAL

- normal dist can be used to approximate the binomial distribution
  (continuous dist) (discrete probability dist)
- solution: round off & approximate through binning appropriately



⊛ +0.5 since continuous

$N\pi$ and $N(1-\pi) > 10$ → OK Approx

6) STATISTICAL LITERACY DEMO

- evaluating tail risk (risk analysis)
  ↳ investment that moves > 3σ from μ

△ Larsen & Marx - 5th ed:

4.3 The Normal Distribution

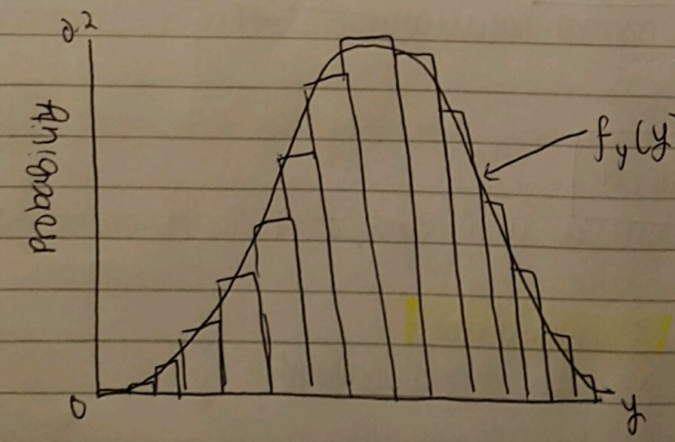
- demoivre:

$$f_3(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

can be used to estimate

$$P\left[a \le \frac{X - n\left(\frac{1}{2}\right)}{\sqrt{n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} \le b\right]$$

where X = binomial random var w/ large n, $p = 1/2$



THEOREM 4.3.1: Let X be a binomial random var defined on n indpt trials for w/c $p = P(\text{success})$. For any #s a and b:

$$\lim_{n\to\infty} P\left[a \le \frac{X - np}{\sqrt{np(1-p)}} \le b\right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, dz$$

- Area under the curve = prob of being within a given region

→ the formula for the normal probability distribution w/ μ and σ.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

← Gaussian curve

- finding areas under the standard normal curve:
  1) using a normal table
  2) algorithm

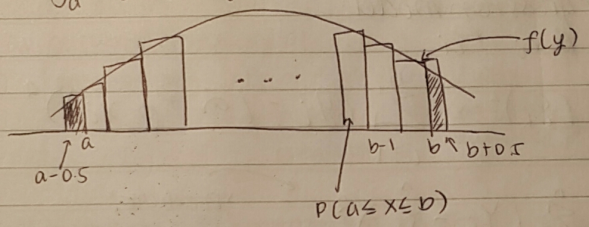
} give the CDF (CUMULATIVE DISTRIBUTION FUNCTION)

$F_Z(z) = P(Z \leq z)$ associated w/ z

& { & compute z score
    & find z score on normal table

- the continuity correction
  → extending $a-0.5$ to $b+0.5$

$$\int_a^b f(y)\,dy \approx P(a \leq X \leq b)$$



- rule of thumb: DeMoivre-Laplace limit should only be used if magnitude of n and p are such that $n > 9\frac{p}{1-p}$ and $n > 9\frac{1-p}{p}$

SUM OF SQUARES

- residual sum of squares - to help you decide if a statistical model is a good fit for your data
  - measures overall diff b/n your data & values predicted by estimation model
  - "Residual" - distance from data point to regression line (measure of)

Total SS = Explained SS + Residual SS

- Total SS = "how much variation is there in the dependent variable"?

$$\text{Total SS} = \sum (y_i - \bar{y})^2$$

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2$$ → Regression.

④

- explained ss = "how much of the variance in the dpt var does our model explain"?

$$\text{explained ss} = \sum (\hat{y} - \bar{y})^2$$

- Residual sum of squares = "how much of the dependent var variance did our model NOT explain"?

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{n} (\varepsilon_i)^2$$

Linear simple Regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$RSS = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

(mean squared error)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

"predictor"

source error

- estimator that measures the average of the squares of the errors
- "risk function"
- measure of quality of an estimator (non-negative)
- closer to 0 = better

- vector of n predictions generated from a sample of n data points
- Y = vector of observed values of var being predicted