7 June 2016

Colors

Black : possible student exercise (4)

Blue : reminders for me (1)

Red : teaching notes (3)

Green : calculations (2)

SS ≡ Synchronous Session

Q ≡ Question

Part 1 (1-7) : JUD ≡ Judgment, Uncertainty, and Decisions

don't deal with samples - much less complicated

Part 2 (8-14) : DAD ≡ Data Analysis and Decisions

Week		
1	SS 1 intro	8
2	SS 2 probabilities	9
3	3 special distributions	10
4	office hours w/me?	11
5	4 regression analysis	12
6	ditto	13
7	SS 5 summary & review?	14 10 summary & review

? front 15 / end 15 — ½ hr Qs, ½ hr directed

- ingredients : a) problem discussion
b) lecture discussion
c) reading discussion → papers

dfc 18 May 2016 . Notes on MBAd 224 videos

Session 1 : Exploring Data ("in depth")

"characterize efficiently" ↳ "Judgement, Uncertainty, & Decisions"

- "essence of a body of data"
- "summary ⓠ" ⇄ data ; but data → decision ; what can I do with these data?

* what decision can I make?

- best way to view data
  - frequency = # time^{-1}
  - "absolute" vs. "relative"

13:36 works 'cause calculating "proportion-years" ; Σ = cumulative

THINK: WHY ? ? ? or HOW

$$\text{proportion} = \frac{\#\text{ people} < x}{\text{total people}}$$

% people ↳ ave yrs person^{-1}

use Ansombe's Quartet? yes ☐

- importance of plotting

- NB : can subtract years from average (in SD calc) 'cause they have the same dimensions, first, and units [years] } * emphasize

* exercise : demonstrate original mean calculation equivalent to professor's method:

A) $\frac{1}{N}\sum_{i=1}^{N}(\Delta t)_i = \mu$

$= \left[1 + 1 + 2 + \cdots 9 + \cdots 10\right]\frac{1}{N}$

$= \frac{1}{N}\left[n_1(\Delta t)_1 + n_2(\Delta t)_2 + \cdots\right]$

$= \frac{n_1}{N}(\Delta t_1) + \cdots$

$= \sum_{i=1}^{10}\frac{N}{N}(\Delta t_i)$

B) $\sigma^2 = \sum_{n=1}^{N}\left[\mu - (\Delta t)_i\right]^2\frac{1}{N}$

$\sum_{i=1}^{10}\left(\mu - f_i\,\Delta t_i\right)$

$10\mu - f_1(\Delta t)_1 - f_2(\Delta t)_2$

$10\mu - \frac{n_1}{N}(\Delta t)_1 -$

$\mu - \frac{n_1}{N}(\Delta t)_1 + \mu - \frac{n_2}{N}(\Delta t)_2 + \cdots$

$\mu = \frac{\sum \Delta t_i}{N}$

$\frac{1}{N}\left((\Delta t) - (\Delta t)_1 + (\Delta t - (\Delta t)_1) \right) + \ldots$

don't see how we can understand skewness (or kurtosis?) w/o calculating : exercise?

What is a "random variable" e.g. in this example, truly random? Q

interquartile range - middle 50%

Qpww: around dots min + max? [ct box + whisker]

IQ ≡ normal?!

"policy implications" → decisions possible

$\frac{\sigma}{\mu}$: coefficient of variation; "relative dispersion"

1) key measures
2) policy implications

Session 2: Probability Concepts

- classical, empirical (repeat), subjective (estimate)

prior knowledge

"on the other hand, judgment is fallible"

1) a) discussion; examples
b) use most?
c) best? "strengths + weaknesses"
d) evaluate

"all ok"

SS: REQUIRE QUESTIONS.

6 June

2) What is meant by "large"? (law of large numbers)
3) experiment: process, repeatable "in which the results are uncertain"
4) event vs simple event (outcome)
(collection) of

X distribution → what we expect : management (neutrino experiment)
→ future

what do we mean by this?

or: "∪" : $P(A \cup B)$ → A or B or (A and B)

and: ∩ → both

complement: $P(\overline{A}) = 1 - P(A)$ "not A"

$P(A|B)$ = conditional probability

→ independent if $P(A|B) = P(A)$

"mutually exclusive": no simple events in common

general addition law: $P(B \text{ or } C) = P(B) + P(C) - P(B \text{ and } C)$ → zero if m.e.

Does everyone understand that writing as a decimal gives (a) probability directly? That "per cent" means "per 100"?

(e.g., in smoking example) → explain chart numbers? (29^m 20^s)

Was "marginal probability" explained? ($\approx 31^m$ 30^s)

Can also go through frequency, i.e., just numbers, first, before division by total

→ 'cause easier to think in terms of counting people (integers?)

Pt: how you think about this can make a difference

↳ and in what order

doesn't really explain this conditional probability example →

$$\frac{55 \text{ non smoke females}}{100 \text{ females}} = \frac{55/250}{100/250}$$

Exercise: create integer table first - then do probabilities

exercise? marginal, conditional, joint

weight lifting → have to do it, e.g. Monte Hall Problem

↳ 5 problems for me!!

1) Monte Hall (1)

Doors: [A prize] B C

switch

i) choose A, open B or C → lose
ii) choose B, open C → win
iii) choose C, open B → win

2/3

2) Bayes Rule: $P(A|X)P(X) = P(X|A)P(A)$

Door A is chosen ⇔ in A

B has been opened (with X goat)

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

$$\equiv \frac{\frac{1}{2}\cdot\frac{1}{3}}{(\frac{1}{2})}$$: opening

prize is in A, B, or C; prob of B

$\frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 = \frac{1}{2}$

do $\frac{1}{2} + 0 + 1$, then divide by 3

Q numerical coincidence?

or are they truly independent?

$P(X|A) = P(X)$

The results are very counterintuitive. For the three-door problem, if the contestant is correct on the first choice, then he or she will be incorrect after a switch. Or, if a contestant was incorrect on the first choice, he or she will be correct after a switch. Since the probability of being incorrect on the first choice is 0.667, then the probability of being correct after a switch is 0.6667.

onlinestatbook.com/2/probability/monty_hall_demo.html

7 June 2016

Session 3: Special Probability Distributions

- factorials
- spreadsheet created for binary calc

⇒ each trial has 1 of only 2 outcomes
- same probability; no hysteresis

fires on car, truck

- point out: 0.9^n gives same result as calculation did it the opposite way ("success" = problem)

$$P(y) = \frac{n!}{y!(n-y)!} s^y (1-s)^{n-y}$$; e.g., $n=3$, $s=\pi$

$y=2$; $\frac{3!}{2!\,1!}\pi^2(1-\pi)$ ✓

$E(y) = ns$

$\sigma^2(y) = ns(1-s)$

as $n \to \infty$, $P \to$ Gaussian (unless s extreme)

an infinite pool of samples, i.e., continual replacement — unlike finite, known number of objects

× Hypergeometric (Hypgeomdist)

↳ depends on history (i.e., no longer have const prob. across trials)

- "sampling without replacement"

$$P(y) = \frac{\binom{N_s}{y}\binom{N_F}{n-y}}{\binom{N}{n}}$$ S = success, F = failure

$$E(y) = \frac{n N_s}{N_{total}}$$

× Geometric "very different"

special case of the Negative Binomial Dist

- for modeling the runs of consecutive failures in repeated independent trials before getting first success

$P(y) = (1-s)^{y-1} s$; $E(y) = 1/s$; $\sigma^2 = \frac{1-s}{s^2}$

Negbinomdist (F, #S, s) in Excel

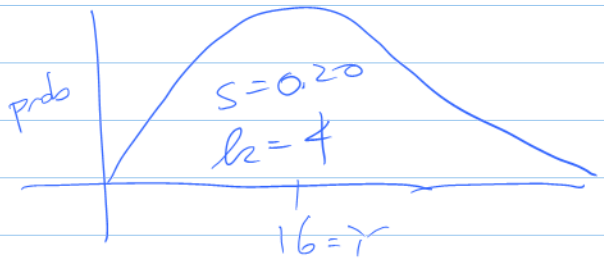
↑ # of failures ↑ number of successes (more general)

55 Discussion of why it might look like that. ; always 1 in geometric dist

number of trials to obtain k successes: $P(y) = \frac{(y-1)!}{(k-1)!\,(y-k)!} s^k (1-s)^{y-k}$ → geo when k=1

Part b: show ↗

$E(Y) = k/s$; $\sigma^2 = \frac{k(1-s)}{s^2}$



x Poisson $P_{(Y)}(y) = \frac{e^{-\mu}\mu^y}{y!}$; μ = mean ; no upper limit (unlike binomial)

x probability tree too complex to draw → Google it? discuss?? SS?

x Poisson (y, μ, cum or not)

three assumptions: 1) exactly one event → prob small + constant
2) two events simultaneously ≈ 0
3) events independent

SS Discussion: What if (2) is violated?

x Exponential Distribution → (distribution of) waiting time before next event

- gives example of toll booth; isn't traffic sinusoidal? makes sense

calculus → not covered

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

what does this mean here? ?? $Y = f(y)$

NORMDIST (x, μ, σ, true → cumulative)

everyone get the difference calculation to get 130 - 100 range?

x

7 distributions reviewed

8 June 2016

Session 4: Decision Making Under Uncertainty

1) set of decisions
2) set of outcomes, w/ probabilities
3) "payoffs"

} key elements in decision making associated w/ decisions

three possible strategies:
a) Maximin: minimizes downside risk
b) Maximax: maximizes upside potential
c) expected value

- point of "unfair coin": you don't know
sensitivity analysis: systematic altering of inputs
but "unfair" coin is one ⇒

× p230f: Decision trees
□ t, decision (decision known)
○ t, probability, but result known
△ end: (no uncertainty conditional on previous events)

check if specific diagram(s) still used in book (no)

"folding back procedure" Q in SS?
↳ inconsistent (? too strong) with

× PWW Q: why 20, 25, 30 on diagram? ? = Δ(win-lost)?
SS Q: as Phil says, make sure understand all numbers on diagram

× "utility" ≈ "usefullness"; risk averse
→ utility maximizers vs EMV maximizers?
× mention Bazerman book

Session 5: Regression Analysis, Time Series Analysis, & Forecasting

"modeling is at the forefront of what managers do"
SS Q? T/F?? Should it be?

ind var (x) → "predictor"
↳ Q: bad terminology? I argue that it causes data confusion and even worse!

dep var (y) → "criterion"

"population regression function" → $E(Y) = \beta_0 + \beta_1 X$

straight line formula: $y = mx + b$

- graphing: good to put zero point

the slope has dimensions!! SS definitely!

and so do x & y

- and label the graph → "sale price/$1000) or "sale price in 1000s of $"

Excel: Intercept (Y range, X range)
Google Slope ( ) ? does LLS

Point out Google S.S. work too (from the beginning) ?

- method of "ordinary" L.S.
(hmm

"subsequent course" → part 2?

QPWW

- hat($\hat{y}$) means predicted
- outliers

correlation vs. regression Q (PWW)
↳ no distinction 'twixt dependent & ind. variable

Excel: Correl (X range, Y range)

$$\hat{r}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{s_x s_y}$$

multiple linear r: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$

SS: Δ quantitative & qualitative; e.g., near a school can be quantitative easily!
[PWW] (vs numeric/nonnumeric)

- additive vs. interactive (non-additive)
  - ↳ slope changes, e.g., w/ quadrant of city
  - ↳ between variables (not people)

x

time series: $x \to t$ and so $y(x) \to y(t)$

log y vs. t ↳ %

$\log y = mt + b$

$y = e^{mt}$

"% change in y = f(t)" → ratio

$\log y_2 - \log y_1 = mt_2 - mt_1 \to m\Delta t = y_2/y_1$

≃ 37:30 ? SS →

(x slide sequence ? this time or always?

mentions tipt good in forecasting models; explore SS examples??

should be on last slide, Alternative Models #19

⇒ switch last two slides

Session 6: Optimization

QSS: What's the difference between "optimization" and numerical solution

- can use Google Sheets, too
- Simplex Method: are of many (?) to do the search
- "constraints"; QSS: LP used in your organization?

PC Tech Optimization problem still in textbook?

- how to avoid VoPP??

- "Key Elements":
  a) input variable
  b) decision variables (changing)
  c) objective function (maximize)
  d) other
  e) constraints { $80 n_{Basics} + $120 n_{XPs}

but Google S.S. too! helps

- algebraic model is an intermediate step QSS? { solution: 560 Basics, 1200 XPs
- graphical solution: i) not clear why plot XP | Basics ii) what if 3 dimensions?

Do in SS?

Q:

derive straight line for assemble hours constraint

$5 x_1 + 6 x_2 \leq 10{,}000$ (x(basis) g(XP))

and testing hour constraints

typo on plot max is $\frac{10{,}000}{6} = 1666.6$

+ answer

$x_2 \leq \frac{10000 - 5x_1}{6}$

$= 1666.6 - \frac{5}{6} x_1$

+ units are ?

WHY IS THAT THE ANSWER?

define isopropyl line

$\partial x = 2000$

$\frac{10{,}000}{6} - \frac{10{,}000}{6} = 0$

QPWW — want to sell more XPs

want to maximize assembly hours, too

want area under line maximized?

↳ "higher"

- Simplex Method for LP
  → says book is excellent here [ ] make one there?!
  + Google SS discussion?
- "quite a lot of time"
- postal problem
- blending models
- logistics models
- aggregate planning models; workforce levels → production schedules
- financial models

↑ should be easy to change to factor model

"comparatively little wiggle room" → inputs highly specified

- mentions 2 more, too hard: integer & non-linear

(but these at least same in principle (I would argue)

probably doesn't use Simplex (but who cares?)

Session 7: Simulation

→ "brief" ; "commercial use" ; not generally Excel
"specialized simulation language"
↳ Python? wasn't listed (later)
"general purpose"
↳ programming + STELLA

→ FC sim 3D ISM SN-dominated AND jet sim.

QSS: Δ "simulation" ≠ "model"?

Wikipedia definition! → (disagree)

Δ optimization → range of values but can also produce values

Not EXCEL!!

- objectives a) model
  b) test vs. empirical data
  c) → future
  d) f(Δ parameters) → "what if"

use of probabilities here, too

- bank/teller example; business, gov, manufacturing w/in e.g.s, ⇒ health-care delivery

SS

if fast → program

- Four advantages: a) no disruption
  b) test before resources
  c) i.d. bottlenecks
  d) insight into (ing factors of system performance

- Discrete vs. Continuous
  ↳ but hydro sims are a hybrid, $\Delta t \neq 0$

- "Model of a System" yes now agreeing w/ my disagreement above
  * "look at the book"

7/14 Jun. 2016 (DATA ANALYZIS & DECISIONS)

Session 8: Sampling + Sampling Distributions

refers to text for more info detail

↳ learn about population from sample data
probability sampling vs judgment sampling
↳ systematic but thus?! convenience but

4 types: simple, systematic, stratified, cluster
mostly in course

vs. say "multi-stage"

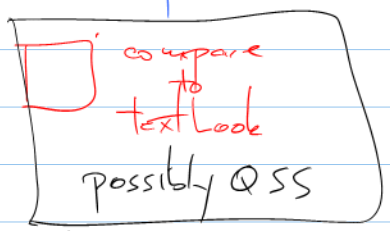
sampling error n.g.
CONCEPT SSQ

pri. obj. ⇒ Capture, at the least cost, the heterogeneity of the population
What PWW calls "characteristics," I would call "goals."
→ efficiency: smallest n/least possible error

Simple Random Sampling (6)
simplicity, enumeration, representativeness, cost, geographical dispersion, efficiency/less sampling error
(often ↔ heterogeneity) (often higher in simple; often low in simple)

5 June 2016
× Gallup, "multi-stage"; but it doesn't use "simple"!

- in course → random sample assumed QSS what about not in course?
- "empirical rule" (from when?): QSS?
  16m00s can change fast-food to E.R.?
- pop parameter vs. sample statistics → sampling error
  (typically not known ← estimate ⇒ estimation process
  - df taken from "dictionary reference dot com" 18m41s
    "the free dictionary"
    18m51s: wikipedia definitions!! oh my



- point vs. interval
  QSS [vs. confidence intervals (frequentist)
  credible intervals (Bayesian)]
- [sampling error vs. non-sampling error] QSS
  QSS: what is truly random?
  → "estimation error" → distribution (prob. dist. of a statistic, e.g., mean, from random sample)
- Central Limit Theorem; remember, mean
  SS: roulette problem ← binary
  Normal estimate ⇒ Δ (ii) as n increases?
  get a better definition! (the book?!)
  "fairly obvious"?!! 1920s?!
  SS (i)
  n ≈ 'Rule' ≈ 30
- if pop = normal, sample mean = normal, ind of n
- $s = \sigma_{sample} \doteq \frac{\sigma_{pop}}{\sqrt{n}}$
  standard error of sample mean =
  "The Empirical Rule" (Applied to Sample Means)
  68% ± σ; 95% ± 2σ; 99.7% ± 3σ
  → error in sample mean, μ (SS) but see Press (SS) w/in ± 2σ
  ≐ 36m40s nice 5 step process (SS); if done 100 times, μ might 95 of

- but then not telling people how numerical answers obtained → make sure no questions

$\sqrt{\frac{N-n}{N-1}}$ = finite population correction
- needed when n/N is not small, i.e., $\geq 5\%$

$\rightarrow \left(\frac{1-n/N}{1-1/N}\right)^{1/2} = \frac{(1-x)^{1/2}}{\sqrt{A}}$, $A = (1-1/N)^{1/2}$, $x = n/N$

Session 9 Confidence Interval Estimation [2]

Announcements

- Ch 8 omissions

check sections text

- Sections from Chapter 8 which you will not be expected to calculate the confidence interval of but will still be expected to know about: confidence interval of, but will still be expected to know about:
  - Confidence interval for a total
  - Confidence interval for a standard deviation
  - Paired samples case of confidence interval for difference in means.
- You will be held responsible on quizzes and the Final Exam for knowing how to calculate the sample size only for the estimate of a single population mean and of a single population proportion

- imp role of α: $\alpha = 1 - (\text{level of confidence})_{\text{decimal}}$

95%: $\bar{x} \pm 2\frac{s}{\sqrt{n}}$ (sample s.d.) → improve → TINV(α, df) (df = n − 1 for mean; degrees of freedom)

t distribution: "The sampling distribution of the standardized sample mean when the sample standard deviation is used in place of the population standard deviation"

Text glossary look defines/explains t! equation [introduce p here too?]

17 June

A note about significant figures, via Bevington: "When quoting the results of an experiment, the number of significant figures given should be approximately one more than that dictated by experimental precision. The reason for including the extra digit is that in computation one significant figure is sometimes lost." (pages 4--5)

So if (in flight example) adding 2-digit integers, quote result $\mu = 18.6$ ✓ correct – e.f. else silly, e.g. upper and lower bounds out to 5 decimals

19th: nice "generic approach to confidence intervals (at least for a mean)

$\bar{x} \pm t\frac{\sigma}{\sqrt{n}}$ (statistic; $\frac{\sigma}{\sqrt{n}}$ standard error; t "certainty")

"sample proportion" ⇒ $\hat{\pi} \pm z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = s$; stressing similarity of form

mug problem: 2 mugs sampled out of 100 → $\hat{\pi} = 0.02$

$t \rightarrow z$ = NORMINV($1-\alpha/2$, 0, 1)] OSC? What is π? 0, 1?

What about difference between 2 population means? ($\bar{x}$)

assuming equal σ^2.

$$\left\{ \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left[\frac{1}{n_1} + \frac{1}{n_2}\right] \right\}^{1/2}$$

actually say "plug + chug"

s^2

= standard error of the difference

$\nu_{tinv} \rightarrow (n_1-1) + (n_2-2) = n_1+n_2-2$

again use TINV, but now called "critical value"

Then, 2 pop proportions ($\hat{\pi}$) difference

$$\left\{ \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2} \right\}^{1/2} = \text{standard error of difference in proportion}$$

- same $z = \text{norminv}(1-\alpha/2, 0, 1)$

Sample size for estimating population mean:

estimate of standard dev = s

$$n \approx \left(\frac{z \cdot \sigma_{est}}{B_T}\right)^2$$

$z \frac{\sigma}{\sqrt{n}} = B$

$z \frac{s}{B} = \sqrt{n}$

Q55

Make students again from desired margin of error derive these two formulas?

For proportion, same z;

and $z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$ (from above) set to B

$\rightarrow B^2 = z^2 \frac{\hat{\pi}(1-\hat{\pi})}{n}$

uses $\hat{\pi}_{1st\ guess} = 0.5$

$\Rightarrow n = \left(\frac{z}{B}\right)^2 \pi(1-\pi)$

'cause you choose a specific value about which to calculate

Section 10 [3] Hypothesis Testing, Part 1

hypotheses refer to the population (μ), not the sample

H_0 = null → disprove, vs. H_A = alternative hypothesis

must always contain an = sign

ok too

try to see if it is disproven

not true that finding alternative true is goal

only confidence, not proof

Q55

$\Rightarrow$ not guilty $\neq$ innocent

("reasonable doubt" = α (= 1 - confidence)

- don't "accept" hypothesis

"Framers of the constitution implicitly understood it." ERIC CIOFFI 17 June 2016

"Null hypothesis are never accepted. We either reject them or fail to reject them." jerry dallal

- directional alternative hypothesis vs non-directional
  - 1 tail → "test of significance approach" (p value)
  - 2 tail → "confidence interval" <> only for non-d.
  - both d + non-d

t_{crit} = TINV $(2\alpha, \nu)$ for directional

↳ significance level changing? (no)

function obscures (α, ν) for non directional

so why 2α vs α?

$$t = \frac{\bar{x} - \mu}{s.e.} \sim = \frac{s}{\sqrt{n}}$$

$\Rightarrow$ is $|t| > t_{crit}$ and direction consistent w/ H_A

vs. $t > t_{crit}$? w/o

"- if you can justify directional" Q ss did he do that?

↳ i.e., more info $\Rightarrow$ more "statistical power"

Says $H_0: \mu \leq 1.4$ "fully equivalent" to $H_0: \mu = 1.4$

(30 min) $H_a: \mu > 1.4$ → directional $H_a: \mu > 1.4$

(referring to textbooks)

Eric likes "operating on that value (1.4)" and justification for directionality is apparent & consistent

now proportion → π and z (text uses p)

z_{crit} = NORMSINV $(1-\alpha, 0, 1)$ directional

$(1 - \alpha/2)$ non directional

$$s.e. = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

π_0 is number you want

$$z = \frac{(\hat{\pi} - \pi_0)}{s.e.}$$; $\hat{\pi}$ you measured (sample)

again, if $|z| > z_{crit}$ and direction is consistent w/ H_A

- h. testing, difference in means; again t

$$s.e. = \left\{ \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} \left[\frac{1}{n_1} + \frac{1}{n_2} \right] \right\}^{1/2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{s.e.}$$ ← the hypothesized difference

- same as before, $|t| > t_{crit}$, direction → H_A { limits; does not work in directional alternative

0) confidence interval approach { "single sample t test"

"LAB REPORT"

Today:
1) t test of a single population mean; proportion
2) independent sample t test, or
3) a t test of two population means

53:50

hypothesis testing approach

58:50

"Things You Should Know How To Do"

"in general the approach is the same"
- give 'em a page or two??

summary each week

Session 11: Advanced Hypothesis Testing [4]

Tough one → make sure have OS

You will be held responsible for the substance, but not the computation of, hypothesis tests of

- equal population variances
- tests of normality

"understanding"

Excel Analysis ToolPak: Analysis Group on Data Tab
→ used in rest of course

~~Statistics~~ Roadmap

Stats
- Descriptive: $\bar{x}$, etc. / μ
- Inferential
  - Estimation
    - Point: estimate the value of a particular parameter in the population
    - Interval
  - Hypoth Testing: whether a particular question is true
    - Previous session and rest of course
    - Populations: ONE | TWO | >2
      - t test
      - difference in μ or proportion
      - "analysis of variance"

⇐ Qss? Explain relationships?

Hyp. Test for Δ in μ / $H_0: \mu_1 = \mu_2$ pt. out?
$H_A: \mu_1 \neq \mu_2$

DO IT → t-test: two sample assuming variance

- e.g., TINV in EXCEL → critical value: "considerable disadvantage"

⇒ p value ≡ conclusion

if $p < \alpha \rightarrow$ reject H_0

otherwise, "fail to reject H_0" ≡ "fail to accept H_a"

p: prob making Type 1 error if reject H_0

(false positive ≡ incorrect rejection

23 June

25th Q55 1.53 vs 1.52 → argue there's reality and there's statistics!] discuss

overheard at an astro conference { "one chance in x statistically ... one chance in a million it's real"

- to make it directional, divide two-tail P by 2 (≡ one-tail [c+])

Ind. Samples vs. Paired/Matched Samples

(2 pops (or more) vs ↑ 1 pop w/ one sample → measure 2 variables

Pepsi vs. Coke example

→ each row represents a different person

"we conclude w/ reasonable certainty"

standing: μ_y same across all populations] null H_0: $\mu_a = \mu_b = \mu_c$

$y \rightarrow$ dependent, continuous variable

μ_y different in at least one pop alternative H_a

⇒ ANOVA ≡ analysis of variance

$\mu_a \neq \mu_b = \mu_c$

or $\mu_a = \mu_b, \mu_a \neq \mu_c$

or $\mu_a = \mu_c, \mu_b \neq \mu_c$

⇒ not $(\mu_a = \mu_b = \mu_c)$

can F test too

various sum of squares - "grand mean"

here, 3 groups

y			
	:	:	:
	:	:	:
i →	A	B	C

1) Total: $\sum_{ij} (y_{ij} - \bar{y})^2$

2) Between: $\sum_{i} n_i (\bar{y}_i - \bar{y})^2$

3) Within: $\sum_{ij} (y_{ij} - \bar{y}_i)^2$

(SS) go over ranges (notation) make them do it?

"test statistic" ⟹ $F = \dfrac{\text{Between} / (I-1)}{\text{Within } (n-I)}$ — # of sets

df (refers to "n − 1") (total # of points)

"a measure of departure from the null hypothesis"

"rejection rule": reject H_0 if $F > F_\alpha$ ("cuts off a right-tail area" "critical F") — need picture (short video?)

- but get answer immediately from P value

so why do we need F test — QSS??

* "Failing to Reject the null hypothesis" (go back earlier when first referenced e.g., "we will never be able to talk how true the null H is")

Review in SS

maybe write them up for earlier SS??

24 June

Session 12: Regression Analysis [5]

- already in JUD, but only w/ total pop; here sample (w/ errors) that we generalize to a population — QSS

see expanded notes, next page

x: ind variable or "predictor" — I don't like this term
y: dep — "criterion" — or this one

Anscombe's Quartet here, quickly again

and get values on line one

remind that that's intercept — maybe in 5-min video

"predicted" values or "expected" values ⟺ $E(y) = \beta_0 + \beta_1 x$ — of $y = mx + b$

"population regression function"

OLS = ordinary least squares (have entire pop)

Δ $b_0 + \beta_0$, $b_1 + \beta_1$ → sampling error

now $b_0 + b_1$ are the estimates, from the sample, of the population

"the Q of interest [often] focus on relationships in the pop," not the sample

i) "specification error" refers to not having all info about what we need to ask to get a valid answer; from before [?]

ii) and sampling error

X Belongs above

26 June

- pww: if only specification error, how well can we predict? add sampling error → can we predict "at all"?

QSS discuss: I argue that in reality, should always be asking "at all" question → "significance" meant "to signify," not "important"

of any "hypothesis" analysis { perhaps check OED

These Qs, at ~ 1Hr in, have been thoroughly discussed at the start

X

- new Q: can we predict dep variable "at all"

Q I've been asking all along!

- can we be sure that $\beta_1 \neq 0$, i.e., is there any relationship??

notation: $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

but $y = \hat{\beta}_0 + \hat{\beta}_1 + \epsilon$ (error term)

→ Hypothesis Testing: $\beta_1 = 0 = H_0$

H_a = "research" hypothesis ↳ alternative like this!

↳ "permits an inference about a pop based on a sample" ⇒ "never if the alternative is true" * only if null can be rejected *

but he's sloppy! (later "true" says)

⇒ $b_0 | \hat{\beta}_0$, $b_1 | \hat{\beta}_1$ "parameter estimates"

vs $\beta_0 | \beta_1$, which indeed are "parameters"

- "standard error of the slope" ≡ $\sqrt{s^2}$ from many slopes; but one works well!

↳ s_{b1}

- "under certain assumptions," remember, distribution of → $\frac{b_1}{s_{b1}}$ → t; dist with n−p deg of freedom (= # of parameters in the model: slope + intercept)

number of obs.

Critical values of $t = b_1 / s_{b1}$ is a value "so extreme that it would be considered very unlikely to occur if the null hypothesis is true."

* compare to earlier hypothesis testing

note figure 3 d.o.f. says ($\pm 3.5^{ish}$)

again $\alpha = 0.05$ → Type 1 Error Rate ⇒ 2.5% on each end

CONVENTIONAL

$t > 3.182$

$t < -3.182$

get st. to draw and verify numbers

"classical" vs p value using

But can I see it better when I think of the t dist and the critical t value from TINV

is the p value less than α? → reject H_0

p value a percentage vs t value which b_1 measured in s.d.

fractional areas under the tails

* but probably earlier, when first encountering p values

[ct] possibly refer st. here first

Two Step Rule for directional alternative hypotheses

$H_0: \beta_1 \leq 0$, $H_a: \beta_1 > 0$

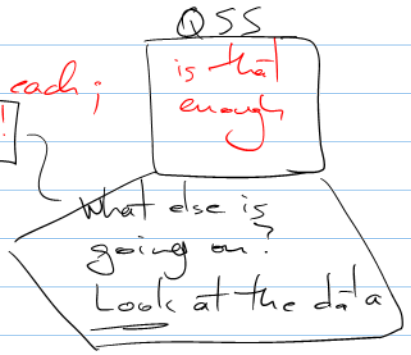
1) divide p value in half
2) directionally consistent?

[d] p vs. t
5 minutes somewhere explained earlier?

Session 13: MuHiple Regression

25 June (vs. in JMP) now

12 topics in intro! in 63 min → ≈5.25 min each; in review example 500ft² → 283K!



samples differ

we don't specify perfectly

Sampling ⊕ specification errors in → $e_i = y_i - \hat{y}_i$

* CONSIDER TOOL PATH from the get-go *

was this ever defined previously?

– residual value

$RMS = \left[\frac{\Sigma e_i^2}{n-2}\right]^{1/2}$ = root mean square error = standard error of the estimate = standard deviation of the residuals

can go to ∞

Coefficient of Determination (R^2) = the proportion of reduction in the error

between 1 and zero

$1 - \frac{\Sigma e_i^2}{\Sigma (y_i - \bar{y})^2} = 1 - \frac{y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2}{y_i^2 - 2y_i\bar{y} + \bar{y}^2}$

26 Jun measuring "goodness of fit"

$= 1 - \frac{1 - 2\frac{\hat{y}_i}{y_i} + \left(\frac{\hat{y}_i}{y_i}\right)^2}{1 - 2\frac{\bar{y}}{y_i} + \left(\frac{\bar{y}}{y_i}\right)^2}$

if all $\hat{y}_i = y_i$ → numerator goes to zero

Multiple Regression

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + e$

partial slope coefficients; no. of parameters

β_i = slope 'tween y and x_i "holding constant" or "controlling for"

plus "additivity" assumption => slope doesn't change all other, $x_{j \neq i}$

e.g., $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

β_1 holds "at all levels of all the other ind. variables"

i.e., find β_1; same as x_2 varies

⇒ β_1 shows how y changes when x_1 changes given that x_2 doesn't changes

BSS?

Unique vs Shared Reduction in Error (Venn Diagrams)

R^2 = Model Coefficient of Determination → shared plus unique gives reduction in error by all x_i's

Squared Semipartial Correlation Coefficient: unique contrib. by each x_i

[not calculated in this course]

N.B. if $b(\beta) \simeq 0$, small unique relationship w/ y but not necessarily fundamentally

$$\text{Root MSE} = \left[\frac{\sum e_i^2}{n-p}\right]^{1/2}$$

Root MSE: entire shaded area (in Venn Diagrams)

$n-p$: number of parameters in model (= in single regression)

Standard Error of the Slope from our sample is s_{bk}

s_{bk} estimates s.d. of b_k on repeated sampling

$\frac{b_k}{s_{bk}} \simeq$ normal → actually t w/ $n-p$ d.o.f.

→ from TINV function

again, critical value of t; typically $\alpha = 0.5$

again, p values and t w/ no good explanation of why they are equivalent

- make sure to do first time t introduced!

again, if $p < \alpha$ (nondirectional!), reject H_0

- if not, offers no unique contribution

Global F Test tests null that all βs $= 0$ ($\equiv R^2 = 0$)

larger F, more likely null to be rejected

"significance F" is p value associated w/ F

x if you want single best β_1, run multiple simple regressions, 'cause multiple regression tells you about unique contribution of each β

x Two Step Conditions, same as before

x all above assumed ADDITIVITY

Session 14: Regression Assumptions [7]

↑ n.b., not multiple, but any

1) Normally Distributed Error → average error is zero
assumed for the population AND the sample
if not, t calc no good

QSS – on all?

2) Homoskedasticity: previously: equal variances in two populations
now, in regression: variances of observations around fitted values are same (i.e., diff x values)

3) Linearity! "frequently unwarranted"
says can test, but difficult
⇒ can 'look!!' (Ansombe again)

Today: "more substantial discussion"

4) Uncorrelated Error Terms; Serial Uncorrelation Assumption
(among different observations)

"limited view"

5) Additivity – slope for x_1 doesn't depend on x_2
(only in multiple regression)

- Data Analysis Course looks at these in detail

here, only relationship between y and x_1, say is the slope
→ that doesn't change (parallel lines)
- pretty much same answer as last week
but now addition ⇒ only 1 p value!

asks: "is it reasonable?" → "in many cases, 'no'!"
⇒ we assume it because it is "convenient" ⇒ "what a silly reason!"

⇒ go to an Interactive Model
⇒ add $x_1 \cdot x_2$ = "the interaction term"
- b, β relationships "much more complicated"
→ now illustrates relationship w/ y when all other x_i = zero!
- include only interaction terms that "make sense" $j \neq 1$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$
$$= \underbrace{(\beta_1 + \beta_{12} x_2)}_{\text{slope when holding } x_2 \text{ fixed}} x_1 + \beta_2 x_2$$

QSS – derive

Why not teach this from the start

back to #4, serially correlated error: $y = 2x + u$ ≡ Actual

if $u = u(x)$: great example at 30m 50s

can get completely mis-leading line } $y = -3 + 3x + u$ ↳ always zero!

⇒ inflated t or F statistic ⇒ s_u^2 = zero!

→ greater probability of Type 1 error

? serial correlation of data ≡ auto correlation

another example (u_t correlated w/ u_{t+1}) : diffuse groupings!

First Order Auto Regressive Process

$y(t) = b_0 + b_1 t + u_t$ $\quad u_t = \rho u_{t-1} + v_t$ (first-order correlation terms; "white noise"; true stochastic error)

⇒ $\langle b_0 \rangle$, $\langle b_1 \rangle$ unaffected by first order correlation but variance is ("sometimes greatly")

$\rho > 0$ more serious (magnification of effect)

→ t value too large } reject H_0 too often (Type 1)

" " small

- assuming first-order : go back one time period to capture auto correlation (all of the effect)

"auto regressive process"

- Conventional Tests for Autocorrelation ? only with time series?

- two basic types : 1. distribution-free tests

a) Geary Test : count sign changes in residuals – not too many, not too few

AGAIN, LOOK!!

b) Durbin-Watson (d-statistic) Test

$$d = \frac{\sum_t (\hat{e}_t - \hat{e}_{t-1})}{\sum \hat{e}_t^2}$$ where $\hat{e}_t$ is OLS regression residual

→ positive autocorrelation, ratio small

negative larger

$\rho = 0 \rightarrow d = 2$

$\rho = -1 \rightarrow d = 4$ no negative

$\rho = 1 \rightarrow d = 0$: strong positive correlation

"Regions of Significance" from 0 to 4

"reject" vs. "uncertainty"

correlation "suspected" $d \leq 1.5$ positive

$d \geq 2.5$ negative

Six-Step Process in Excel!

check google sheets! ✓

1951 papers

$|r|\,|d|$ = SUMXMY2 (C2:C50, C1:C49) / SUMSQ (C1:C50)

- if n.g., Excel n.g.! → SAS

X "Retrospective Overview of the Course"