6/6/11

Lecture #9

Admin

- June 8,9 no class
  talk about holiday
- Midterm review session
  8-10pm here
- office Hrs today & tomorrow
- 2e,f unassigned

Plan

- Log transform + demo
- Sampling ch13
- Cont. r.v.'s Midterm goes

p188-188 Log transform (do demo) same as p184!

Height	ln(H)		
64"	4.159		
68"	4.220		
70"	4.249		
71"	4.263		
81"	4.394		
$\bar{X} = 70.8$	$\bar{Y} = 4.257$	$e^{\bar{Y}} = 70.6''$	✗
med = 70"	med = 4.249	$e^{med} = 70''$	✓
80%ile = 71"	80%ile = 4.263	$e^{80\%ile} = 71''$	✓

strictly monotonic, 1:1

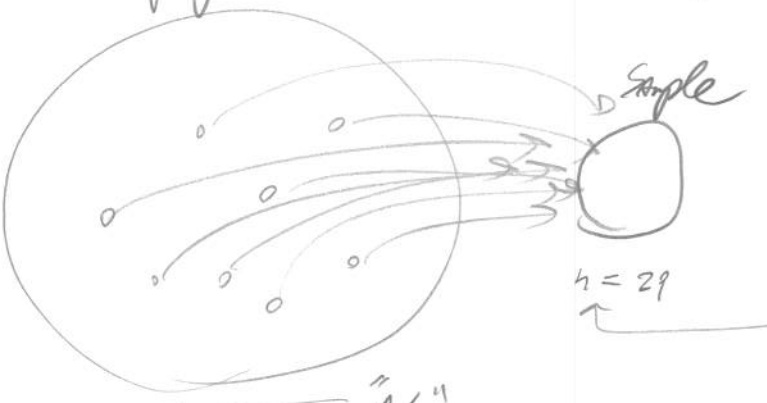
log is a "quantile-preserving" transformation

Why log transform? Reasons we'll get to tomorrow.

Ch 13: Sample / Surveys

Data: realizations of r.v.'s. Where do those r.v.'s come from? How should we "harvest" the realizations? This is called "sampling".

"collect data"



IMP screws things up

$N = \infty$

for all examples in this class. We're going to forget about small N

Population: all possible items, people, bears, or proper

Sample: the subset we collect data about.

In general, a "representative" sample reflects the population. A non-representative sample is "biased".

representative:

biased:

Randomization: pick members from population at random: suffices

No matter size of pop., sample does not need to change.

Sometimes it's easy.

Bank Account #1
Bank Acct #2
⋮
Bank Acct #N

→ #173, #052, #2193, #1057 } $n = 4$

mimics on avg the population.

Harder: sample from all unemployed people (macro)

Who is unemployed? Definitions... Need to study

Randomization: representative, no bias... we can use sample statistics from the sample to estimate pop parameters of the population, e.g. → book says $\bar{Y}$ ... not sure why..

$\bar{X}$ estimates μ
S^2 estimates σ^2

Inference

p311
ignore r, ρ, b, β

Most powerful tool in statistics.

⇒ much better to have small unbiased sample than large biased sample
it's not the "n" that matters... it's how you got the n.
But, once you did it right, the more n, the more accurate

Randomization feasibility:

$Pop = \{ X_1, X_2, \ldots, X_N \}$

$Smp \subset Pop$ s.t. $|Smp| = n$

If each sample is equally likely, then it is considered a "Simple random sample" (SRS), the gold standard.

How to get an SRS? Sampling frame: the list of items in the Pop.
Randomization: take random #'s and sort

(1)

Pop	rand #
X_1	0.43
X_2	0.87
X_3	0.11
⋮	
X_N	0.59

← # btwn 0 and 1

(2) Order ..., take first n

Systematic sampling: get every 10th person on an alphabetical list (if you don't have a computer or it's in print)

You have to make sure name is not correlated with whatever you measure.

Sampling Variation

We flipped 10 coins. We expect 5 to be H, but it could have been 4 H or 6 H or 3 H or 7 H...

Why? Data is random because it comes from r.v.'s.

Once we get to Hyp. Testing, we must know we are making decisions from random data — like!

→ Every SRS will have different Sample statistics.

$\bar{X}$: the r.v., $\bar{x}$: the realization for the random density huge variation

From now on, for everything else we use in the class, we assume SRS.... If we don't have an SRS, everything we do is WRONG.

More Complex methods

① Stratified Random sample: Split sampling frame into homogenous groups called strata. p.318 has really good example.

Imagine hotel: 99% tourists 1% business. If you do an SRS, you won't learn anything about the business guests, so you take a larger sample from those strata. This is not an SRS but can be adjusted if you want to account for this...

Overrepresented? Who?

A type of Stratified sampling is Cluster Sampling: the geographical units e.g. census tracts are the strata.

ex 13.2 is a good example of inference situation

Census: Sample all N... really impractical, expensive

How to clean up a sample?

a) Voluntary response: send a bunch of surveys out... whoever sends them back call us SRS. Problem?

b) Convenience Sampling: ask whoever is around...

c) Survivorship bias: sometimes people in some groups live too long or die too early to make judgments about a parameter

---

What's on the midterm?

Set theory: unusual applications...

Counting: perm, comb methods

Prob: definition, axioms, rules, applications, conditional prob, Bayes Rule
trees, ind/dep

ch 1-4 (not p67)
ch 7.2 - ch 9
ch 10.1, 3, 5, 11
ch 13

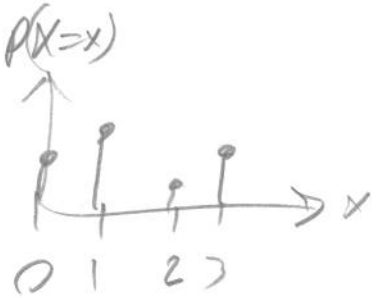
r.v.'s: exp, var, SD, applications, sums, constants, Bernoulli, Binomial, Poisson,

data: basic stats, visualization, distributions, displays

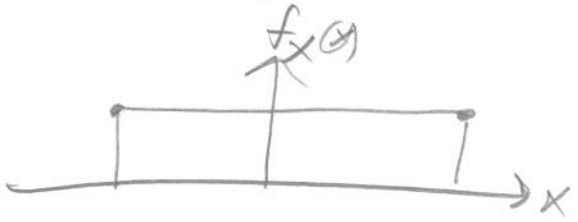
Discrete r.v.'s: outcomes: finite or ctbly finite # of choices, $\{0,1\}$, $\{0,1,2,3,4,5\}$, $\{0,1,2,\ldots\}$, $\{-1, 1, \ldots\}$ etc

Cont r.v.'s: all real / decimal / fractional values acceptable.

→ have PMFs:



→ have density functions:



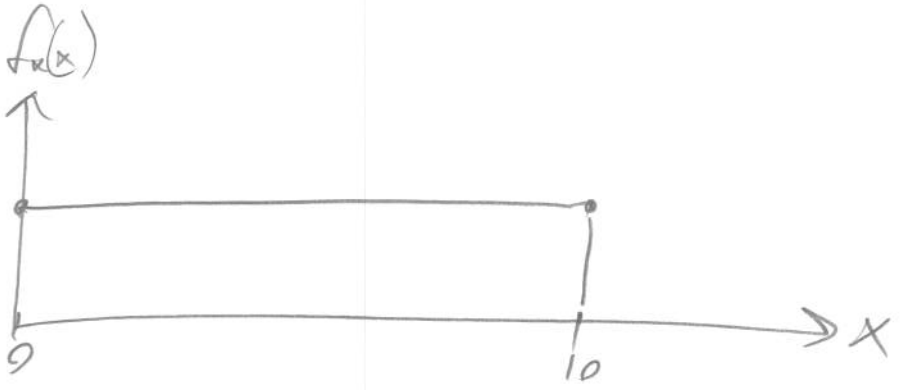
to get probs out of PMFs you sum:

$$P(1 \le X \le 4) = \sum_{i=1}^{4} P(X = i)$$

to get probs out of PDFs you integrate

$$P(1 \le X \le 4) = \int_{1}^{4} f_X(x)\, dx$$

For eagle... bus wait 10 min



so-called uniform density $X \sim U(0,10)$

What is prob you must have 2 and 7 min?

$$P(2 \le X \le 7) = \int_2^7 (.1)\, dx = .1[X]_2^7 = .1(7-2) = \boxed{\frac{1}{2}}$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx, \quad Var[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx \quad \text{weighs}$$

Most special cont. density:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

bell curve