Session 3: Numbers

Li (Sherlly) Xie

Session 3 Flow

Session 2 & 3 Tasks

Break

Terminology

- Population, Sample, Observation
- Probability vs Empirical Distribution

Descriptive Statistics

- sample size
- central tendency measures: mean & median
- variability measures: variance, standard deviation, standard error, range

A Glimpse at the Data

	A	B	C	D	E	F	G	H	I	J	K
1	ID	gender	race	FHDM	history of GERD	gestation	why referred	age	Ht.m	Wt.kg	BMI
2	1	F	Asian	?	no	33	abnormal QR in EKG	16	1.2	46	31.94
3	2	M	asian	no	no	32	had chest pain	12	1	41	41.00
4	3	M	Caucasian	type 1		37	cp, syncope	12	1.54	36	15.18
5	4	F	African american	1		38	cp, abnormal ekg	15	1.6	37	14.45
6	5	F	caucasian	2		38	couplet	14	1.28	41	25.02
7	6	F	africanamerican	father		36	cp	8	0.5	28	112.00
8	7	F	AA	cousin	yes	34	fam hist VT	13	1.1	30	24.79
9	8	M	Cauc	type 2, mother	yes	32	ADHD	12	0.9	30	37.04
10	9	M	A	yes		27	SV beat, cp	16	1.64	36	13.38
11	10	F	Asian	yes		30	syncope	16	1.66	41	14.88
12	11	M	Caucasian		yes	30	cp	11	1.2	30	20.83
13	12	M	Cauc			35	cp	10	1.2	28	19.44
14	13	M	cauc			36	autism	16	1.72	37	12.51
15	14	M	African american	no	yes	32	syncope	8		24	
16	15	M	Caucasian	no	yes	25		13	1.64	33	12.27
17	16	F	Caucasian	yes		41	cp, syncope	17	1.8	48	14.81
18	17	F	Latino		yes	27	syncope	15	1.4	46	23.47
19	18	F	africanamerican			46		15	1.33	46	26.00
20	19	M	africanamerican	uncle		25	on BB for PVC	16	1.12	48	38.27
21	20	M	african Amer		yes	22		12	1.21	40	27.32
22	21	F	Caucasian	yes		36		13	1.26	40	25.20
23	22	M	caucasian			36	ADHD	11	1.04	26	24.04

Session 2 Tasks

1. Convert the qualitative information in "FHDM" and "why referred" into quantitative information.
2. How would you deal with the empty cells in "history of GERD, "why referred", "Ht.m" and "BMI"? Explain your reasoning.
3. Generate 5 statistically testable hypotheses
4. Design a study for 1 of the hypotheses in #3, define the nature of your study.

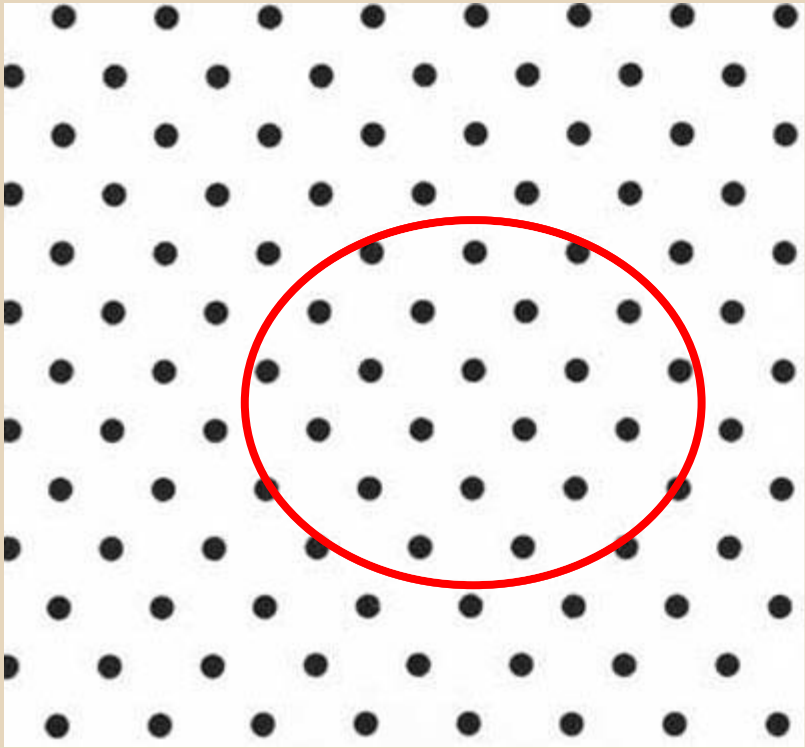
Missing Data

Rule of thumb: if 10% or more observations are missing for a variable in a sample, then that variable is "in danger".

Under 10%: Report the percentage, do complete data analysis, assume the missing observations are missing at random

Example: a data set consists of subjects 1-6, variables A amd B. Subject 1-3 miss variable A, subjects 4-6 miss variable B, removing all missing data leaves NO subjects for analyses for BOTH A and B

Population, Sample, Observation



The Mathematical assumptions

The assumptions we CANNOT change: independent and identically distributed random variables

vs

What we usually get

And randomization does **NOT** save us from this

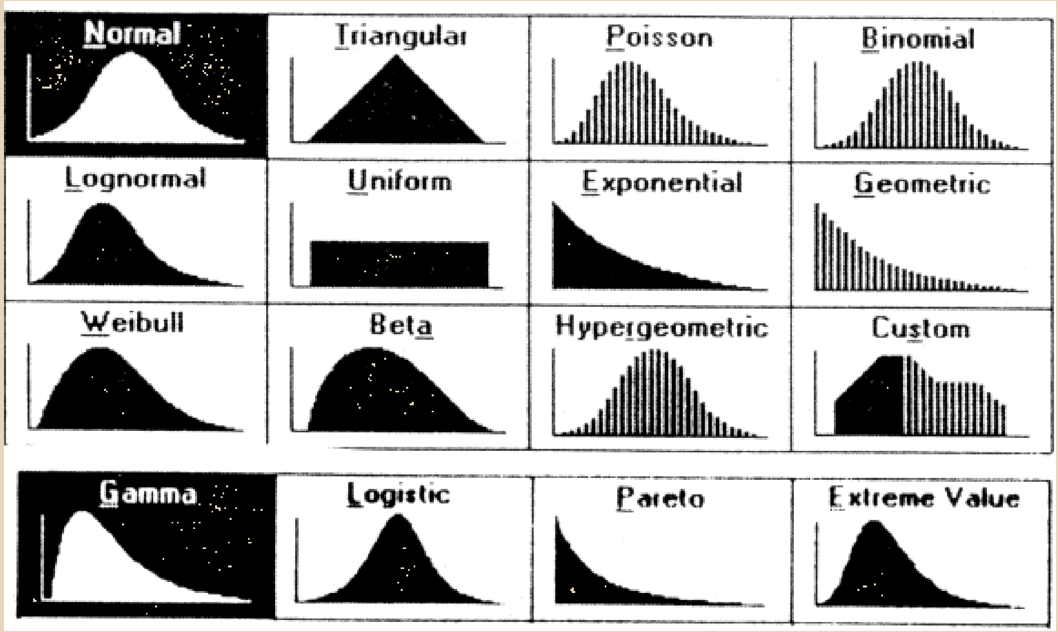
So what does randomization do?

It saves the investigator from the investigator's bias in assigning treatments to the subjects. It does this AND ONLY this.

Randomization does not make the sample representative; It does not give favorable p-values; It does not guarantee "balance" to the placebo vs trt grps; it is a untestable/scientifically AND mathematically unverifiable belief/claim.

Probability vs Empirical Distributions

Usually, it is assumed that random variables has some probability distribution BEFORE the experiment and an empirical (i.e. data) distribution is obtained AFTER the experiment is performed.



Probability distributions are CONVERTIBLE

[probability distributions conversion]

Sample size

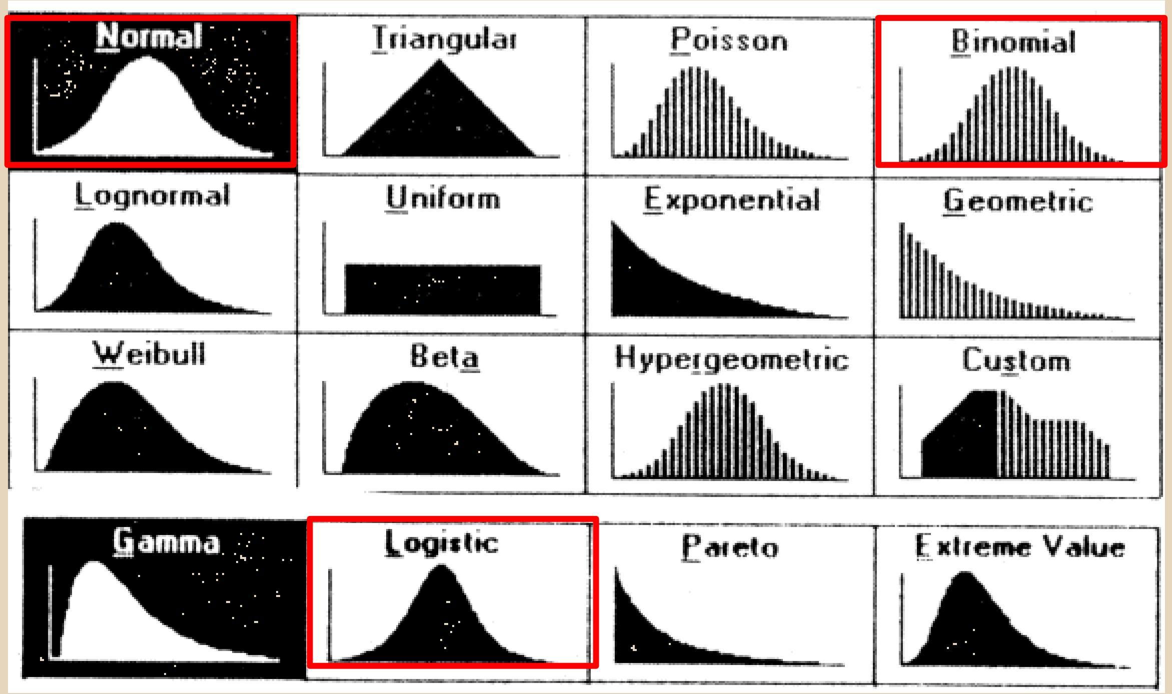
Sample size is the number of observations sampled from the population. The larger the better (the largest sample is the size of the population)

Probability does not apply to statistical inferences made using the entire population

Law of Large Numbers in a simple example:

2 vs 2 billion flips of a fair coin. more samples=closer to "truth"

2 Measures of central tendency: Mean & Median



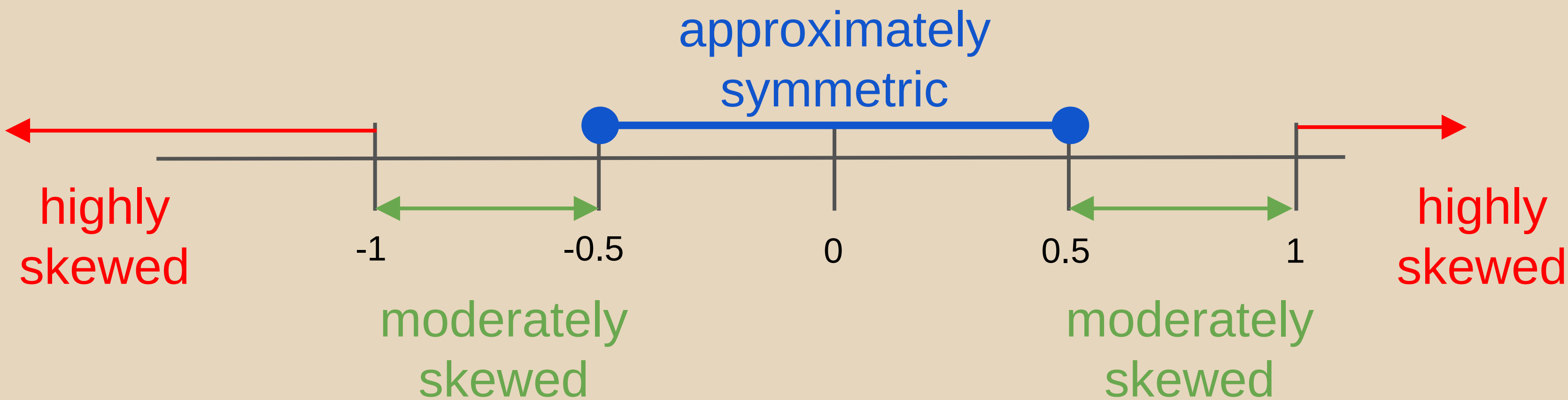
T test, z test and ANOVA has the IMPLICIT assumption that the distribution is normal (at least roughly symmetric)

This is important:

For symmetric distributions, mean=median. For skewed distributions, they are not equal. Median is more "stable" (we call it "robust to outliers") than the mean.

WHICH MEANS MUST CHECK THE DISTRIBUTION before running any tests; Violation of hypothesis test's assumptions weakens the test result

How skewed is skewed?



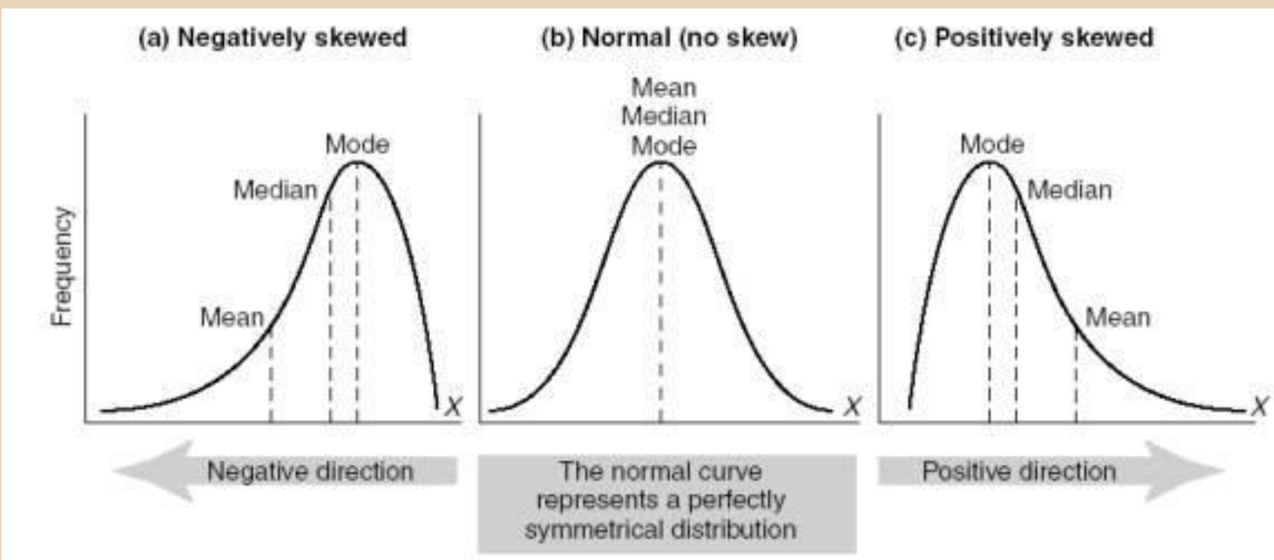


FIGURE 15.6 Examples of normal and skewed distributions

Bulmer, M. G., Principles of Statistics (Dover, 1979)

Mean & Central Limit Theorem

Central Limit Theorem

REGARDLESS OF THE DISTRIBUTION, means of samples (from the same distribution) follow a normal distribution, symmetrically distributed around the "true" mean of the population.

Variability

Variability measures like variance & standard deviation (SD) expresses how far the individual data points are away from the mean

Example

1 sample containing 5 observations: (1,4,3,6,11)

mean=(1+4+3+6+11)/5 = 5

median: middle value of (1,3,4,6,11) = 4

Is this distribution skewed, symmetric or normal?

$$\text{Variance} = \frac{(1-5)^2+(4-5)^2+(3-5)^2+(6-5)^2(11-5)^2}{(5-1)} = 14.5$$

SD = square root of 14.5, ~ 3.8

Range = max - min = 11-1 = 10

standard deviation=square root of variance

standard error=standard deviation of MEANS between samples

What could be inferred about standard error if under repeated sampling, the averages from different samples do not vary much?

Standard error must be small.

What could be inferred about standard error from knowing the value of standard deviation? (assignment problem)

Excel and SPSS Commands: Excel

For this week's task (Due next Tuesday 9am),please explore and use the following Excel commands:

=skew()

=stdev()

=average()

=median()

=max()-min() gives range

=var()

=sqrt(var()) Should give the same results as =stdev()

Excel and SPSS Commands: SPSS

2 commands:

"descriptives"

"frequency"

This week's task will be up on the web shortly.

Next week's main topic:

Graphic exploration & display of data

See you!