Introduction to Statistics

I. What are statistics?

Statistics deals with variation and attempts to draw conclusions from data despite variation.

Trt	Mass	Hem		Trt	Mass	Hem
F	67.6	46.13		C	59.34	36.2
F	71.23	44.23		C	70.74	36.92
F	70.7	46.1		C	72.54	38.96
F	73.6	47.2		C	66.7	45.55
F	76.78	42.53		C	67.5	42.45
F	67.28	39.9		C	65.23	34.07
F	68.6	41.3		C	70.3	43.5
F	68.16	39.48		C	69.75	30.95
F	71.95	46.33		C	64.48	32.76
F	70.65	42.1		C	59.35	36.6
F	63.68	52.9		C	61.1	45.01
F	76	41.33		C	70.53	43.9
F	79.73	43		C	60.58	35.78
F	66.85	41.5		C	67.37	36.57
F	70.08	41.47		C	69.7	40.88
				C	73.18	34.52

A. Definitions

- Data – information pertinent to answering some question
- Population – group to which you are trying to generalize
  - Observational (wing lengths of House Flies)
  - Experimental (wing lengths of males on standard diet)
- Samples – the proportion of population that is measured

Definitions (cont.)

- Experimental Unit – the “thing” that is measured; the smallest unit that is independent of other units and to which we can randomly assign a treatment.
- Random Sample – sample drawn so that all members of a population have equal and independent chance of being included in sample.

B. Roles of Statistics

2 major roles

1. Condense variable information into a summary to convey information (descriptive stats)
2. Assess whether given variability in data are consistent with your hypothesis (inferential stats)

II. Descriptive Statistics

A. Location – where on a scale do the data fall

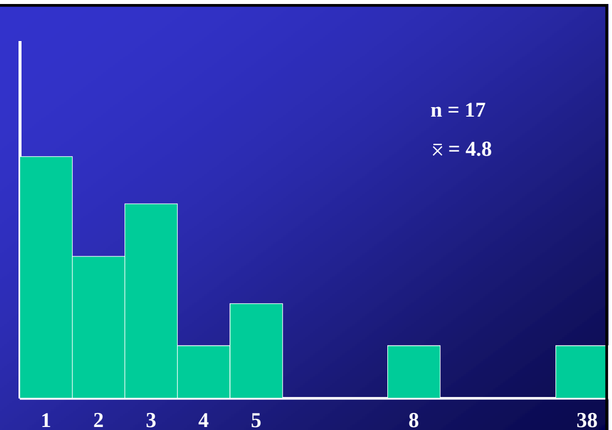
1. Mean – the average of a sample

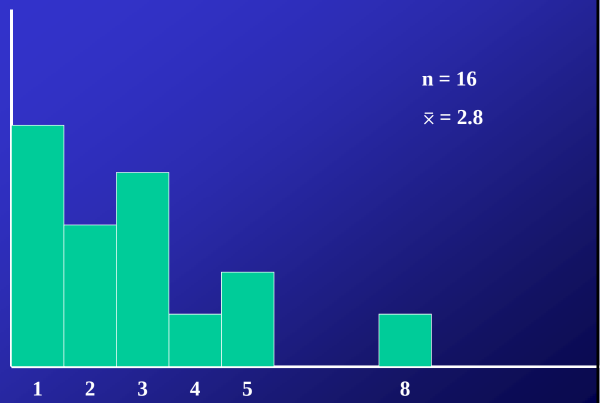
$$\bar{x} = \frac{\sum x}{n}$$

Advantage – simple to compute and interpret

Disadvantage – heavily influenced by extremes

If data are skewed then not good measure





2. Median – middle value, 50% less than and 50% more than

Rank data from smallest to largest – median is rank n+1/2

Odd

14 17 **18** 20 21

Even

14 17 ! 18 20

3. Mode – most frequent value, commonest

Used very infrequently, mostly by Ornithologists

6
5
2
6
7
6
3
6
7
5

B. Dispersion

Spread of data around a central location

1. Range – difference between max. and min., very sensitive to extreme values (same units as original data)
2. Standard Deviation – measure of mean deviation of observations from the mean of the distribution (same units as original data)

(mean distance from the mean)

s = $\sigma = \sqrt{\frac{\sum(\bar{x} - x)^2}{n - 1}}$

3. Variance – quantifies how far each observation is from mean. No units associated with variance. Average of the squared deviations

Var = s^2 =

$$Variance = \sum \frac{(x - \bar{x})^2}{n}$$

Important measure in statistics.

X	*f*	$x = X - \bar{x}$	x^2	fx^2
3.4	2	-0.6	0.36	0.72
3.7	8	-0.3	0.09	0.72
4.0	5	0	0	0
4.3	8	0.3	0.09	0.72
4.6	2	0.6	0.36	0.72
$\bar{x} = 4.0$	n = 25			$\sum$ 2.88

$s^2 = \sum fx^2/n = 2.88/25 = 0.115$ (average of squared deviations)

$s = \sqrt{0.115} = 0.3394$

4. Standard Error – often used synonymously with standard deviation, standard deviation of mean

$se = \sqrt{s^2 / n}$

5. Coefficient of Variation (CV)– std dev expressed as % of mean.

When populations differ (considerably) in means direct comparisons of variance or std deviations not useful.

e.g. larger organisms vary more in size than smaller ones (std dev of elephant tails will be greater than std dev of mouse tails)

CV - compares relative amounts of variation in populations with different means

$$CV = s \cdot 100 / \bar{x}$$

III. Inferential Statistics

A. Hypothesis testing

Goal is to determine if 2 samples differ (were samples drawn from same population).

e.g.

- 2 independent samples drawn from same population
- Calculated means estimated from same population
- Differences result of chance / sampling error

Null hypothesis – means of 2 populations are equal

Alternative hypothesis – means of 2 populations are different.

Test allows us to say with some level of certainty (probability) if we can reject the null and accept the alternative.

Need to determine what magnitude of error we are willing to live with.

	Null hypothesis	
Null	**Accepted**	**Rejected**
True	Correct	Type I error
False	Type II error	Correct

Level of probability traditionally chosen to be 0.05

You are willing to take a 5% chance of rejecting the null when it is in fact true.

Type I error = α

Tests calculate p value for you, but you must determine α BEFORE the experiment.

B. T – test

Ho – $\bar{x}_1 = \bar{x}_2$; means drawn from same population

Ha – $\bar{x}_1 \neq \bar{x}_2$; means differ

Traditional test to determine if means from two samples are different from one another.

1. Assumptions of test

- Individuals sampled randomly from population
- Variances from each sample are not significantly different
- Data are normally distributed

2. Procedures

SPSS