

● ● ● | PSY 464
Advanced Experimental Design

Describing and Exploring Data
The Normal Distribution

1

● ● ● | Overview/Outline

- Questions-problems?
- Exploring/Describing data
 - Organizing/summarizing data
 - Graphical presentations
 - Histogram, Stem and leaf plot, & Boxplot
 - Describing Distributions
 - Central tendency
 - Variability
- Normal Distribution
 - Description, Z-scores, Areas & Probabilities

2

● ● ● | Moving beyond raw data

- Unorganized/interpretable
- Imposing organization
- Ordering
- N = total number of observations (scores)
- f = frequency of each score

3

● ● ● | Y-DACL Scores – File order

0	13	0	4	7	2	0	15	3	6	1	2	8	3	15
6	4	5	16	5	10	4	1	1	6	8	4	4	6	12
4	8	21	6	8	6	6	6	4	3	8	3	19	12	4
6	0	1	8	4	7	6	5	1	7	12	2	4	11	6
5	5	1	5	7	5	8	7	4	8	6	5	8	10	15
9	8	7	4	0	1	1	3	0	13	0	8	10	7	13
7	2	4	14	4	6	15	6							

4

● ● ● | Y-DACL Scores – Ordered

0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
2	2	2	2	3	3	3	3	3	4	4	4	4	4	4
4	4	4	4	4	4	4	4	5	5	5	5	5	5	5
5	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	8	8	8	8	8	8	8
8	8	8	8	9	10	10	10	11	12	12	12	13	13	13
14	15	15	15	15	16	19	21							

5

● ● ● | Simple Frequency Table

Y-DACL	f	Y-DACL	f
0	7	12	3
1	8	13	3
2	4	14	1
3	5	15	4
4	14	16	1
5	8	17	0
6	14	18	0
7	8	19	1
8	11	20	0
9	1	21	1
10	3	22	0
11	1	missing	2
		n=	100

6



Grouped Frequency Table

Y-DACL interval	midpoint	f	cumulative f
0-2	1	19	19
3-5	4	27	46
6-8	7	33	79
9-11	10	5	84
12-14	13	7	91
15-17	16	5	96
18-20	19	1	97
21-23	22	1	98

7



Frequency table from SPSS

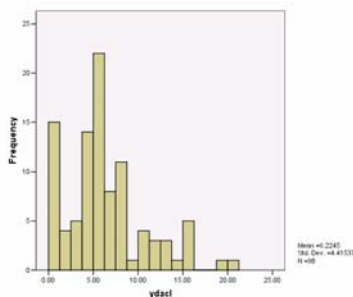
		ydacl			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	7	7.0	7.1	7.1
	1.00	8	8.0	8.2	15.3
	2.00	4	4.0	4.1	19.4
	3.00	5	5.0	5.1	24.5
	4.00	14	14.0	14.3	38.8
	5.00	8	8.0	8.2	46.9
	6.00	14	14.0	14.3	61.2
	7.00	8	8.0	8.2	69.4
	8.00	11	11.0	11.2	80.6
	9.00	1	1.0	1.0	81.6
	10.00	3	3.0	3.1	84.7
	11.00	1	1.0	1.0	85.7
	12.00	3	3.0	3.1	88.8
	13.00	3	3.0	3.1	91.8
	14.00	1	1.0	1.0	92.9
	15.00	4	4.0	4.1	96.9
	16.00	1	1.0	1.0	98.0
	19.00	1	1.0	1.0	99.0
	21.00	1	1.0	1.0	100.0
	Total	98	98.0	100.0	
Missing	System	2	2.0		
	Total	100	100.0		

8



Histogram

○ appropriate for quantitative data



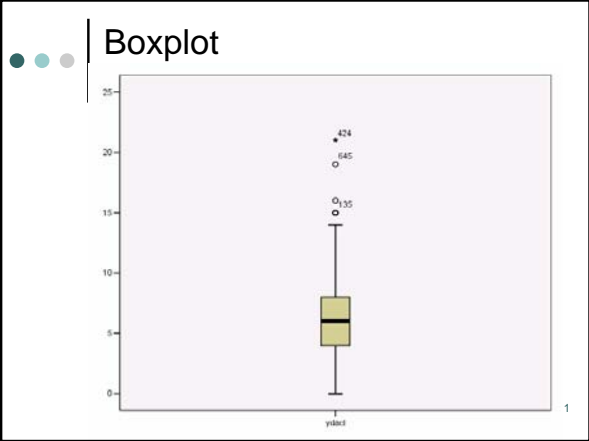
9

Stem and Leaf Plot

- Little or no loss of information

Freq	Stem	Leaf
7.00	0 .	0000000
8.00	1 .	00000000
4.00	2 .	0000
5.00	3 .	00000
14.00	4 .	00000000000000
8.00	5 .	00000000
14.00	6 .	00000000000000
8.00	7 .	00000000
11.00	8 .	0000000000
1.00	9 .	0
3.00	10 .	000
1.00	11 .	0
3.00	12 .	000
3.00	13 .	000

8.00 Extremes (>=14.0)
 Stem width: 1.00
 Each leaf: 1 case(s)



Describing distributions

- Normal distribution: Bell shaped curve
 - most observations concentrated in middle
 - very well-known properties.
- Skewed distributions
 - distribution is not symmetrical
 - tail trails off in one direction or the other
 - greatest frequency of observations not in middle
 - skewness is in which direction?
- Bimodal
- Kurtosis

Measures of central tendency

- o mean
- o median
- o mode

13

Mean

- o a statistic calculated from a sample
- o corresponding population parameter is μ
- o population parameter, we know the exact value with certainty
- o statistic uncertainty is involved
- o \bar{X} is the best estimator of μ

14

Summation notation

- o Σ uppercase Greek sigma
- o tells us to add up an entire group of numbers
- o ΣX means add up all the Xs
- o Page 32-33 Summation Notation
- o $(\Sigma X)^2$ vs. ΣX^2

ΣX	X	Y
ΣX^2	1	2
$(\Sigma X)^2$	3	0
ΣXY	0	3
$\Sigma(X - Y)$	2	4
$(\Sigma X)(\Sigma Y)$	4	2
$\Sigma(X - Y)^2$		

15



Formula for the mean

- where:
- \bar{X} = the mean
- ΣX = add up all the X values
- N = number of scores

$$\bar{X} = \frac{\Sigma X}{N}$$

16



Median

- point at the exact middle of the set of scores.
- list all scores in numerical order, and then locate the point in the center of the sample.
- Median = location $(N+1)/2$
- simplest interpretation it is the score or value where half are higher and half lower
- 50th percentile of a set of scores

17



Mode

- simply the most frequently occurring value in a set of scores
- when giving a modal value should also give an idea of how often it occurred
- can be more than one mode
- if multiple modes, simply list all

18



Which should I use?

- scale of measurement dictates measure of central tendency
- mean with interval/ratio level data, not ordinal/nominal.
- median with ordinal or higher, not nominal.
- mode with any level

19



Skewness

- If the distribution is normal (i.e., bell-shaped), the mean, median and mode are all about equal
- positive skew: Mean > Median > Mode
- negative skew: Mean < Median < Mode
- means from highly skewed distributions can be misleading

20



Variability

- Dispersion around the middle of the distribution, generally the mean
 - Range
 - Interquartile range
 - Variance
 - Standard deviation

21



Range and IQR

- o Range
 - distance between the highest and lowest scores
 - completely dictated by extreme values
- o Interquartile range
 - distance between the 25th & 75th percentile
 - completely ignores extreme values

22



Variance

- o If we subtract each value from the mean and take their average, it always comes out to zero, values above and below the mean cancel each other out
- o Squaring each of these differences eliminates the problem of summing to zero
- o The average squared deviation is the variance

$$\frac{\Sigma(X - \bar{X})^2}{n - 1}$$

23



Standard Deviation

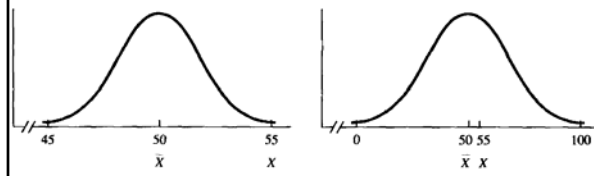
- o Variance provides us with average squared deviation from the mean
- o Taking square root, essentially gets us back to our original measurement scale
- o Definitional form
- o Computational form.

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} \quad \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n - 1}}$$



Standard Scores

- o If we want to directly compare standing on measures with different scales





Standard Scores

- o Even with same mean, distributions may be very different
- o standard scores: convert to common scale
- o z-scores: standard scores expressed in standard deviation units

$$z = \frac{X - \bar{X}}{S}$$

29



A problem

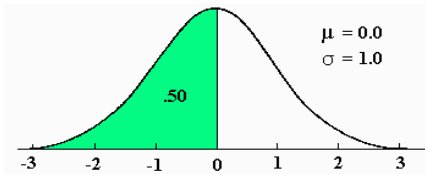
- o **Lets compare two cases in terms of relative standing on level of depression**
- o **A tricky twist – missing data**

idscana	rads	ydacl
151	2.41	-
1248	-	8.00

30

Normal Curve (cont.)

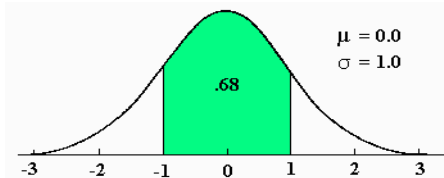
- total area below 0.0 is .50
- symmetrical about the mean, thus area above 0.0 is .50
- generalizes to all normal curves: total area below the value of m is .50 on any normal curve



31

Normal Curve (cont.)

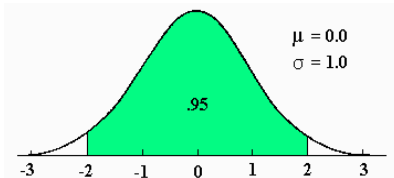
- area between Z-scores of -1.00 and +1.00. It is .68 or 68%
- total area between plus and minus one sigma unit (1 s.d.) on any normal curve is also .68.



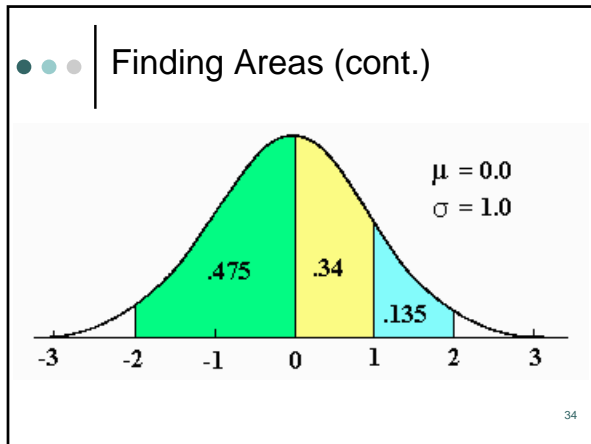
32

Normal Curve (cont.)

- area between Z-scores of -1.96 and +1.96 (about 2) is .95 or 95%
- area (.95) generalizes to plus and minus two sigma units on any normal curve



33



- ### Areas under Normal Curve
- area below a Z-score of 1.0?
 - computed by adding .34 and .50 to get .84
 - area above a Z-score of 1.0?
 - subtract the area just obtained from the total area under the distribution (1.00)
 - $1.00 - .84$ or .16 or 16%
- 35

- ### Areas under Normal Curve
- area between -2.0 and -1.0?
 - first, the area between 0.0 and -2.0 is 1/2 of .95 or .475
 - the .475 includes too much area, the area between 0.0 and -1.0 (.34) must be subtracted from this
 - so, $.475 - .34$ or .135
- 36



Using normal curve

- o Mean = 10, SD = 2
 - what proportion of scores is between 7.5 & 12.5?
 - what proportion of scores is between 7.5 & 10.5?
 - What score separates the lower 40% from the upper 60%?
 - If there were 250 members of population, how many would be expected to score 11 or more?
 - What proportion would be expected to score 9 or more?
 - What score separates the top 10% of scores from the rest?

37
