Variance vs. Variation

Recoding vs Recording

SOC4

Admin Inference Samples Curves Percentiles Z Scores Examples Scales

Where we are…

#	Date	Read (5th)	Due	Area	Lecture Topic	Lab #	Lab Assignment	T Lab	R Lab
1	Tue Aug 26			Orientation	Welcome & Orientation	-			
2	Thu Aug 28	1.1 to 1.4			Basic Terms	-		-	-
3	Tue Sep 2	2.1 & 2.5			Measurement Issues	1	Levels / Age 3x	1	-
4	Thu Sep 4	3.1			Data Reduction	2	Heaven & Hell		1, 2
5	Tue Sep 9	-	HW1	Description	Display & Analysis (Shapes & SPSS)	3	Total Miles	2, 3	
6	Thu Sep 11	3.2 & 3.5			Central Tendency	4	CT		3, 4
7	Tue Sep 16	3.3 & 3.4			Dispersion	5	Dispersion	4, 5	
8	Thu Sep 18	3.7			[illegible]	6	Music index		5, 6
9	Tue Sep 23	4.2		Inference	Probability & Z Scores	7	Standardizing Scores	6, 7	
10	Thu Sep 25	4.3	HW2		Zs & Ps	8	Table A		7, 8
11	Tue Sep 30	3.6 & 5.1			Parameters & Pt Estimation	9	Distances	8, 9	
12	Thu Oct 2	2.2 to 2.4 & 4.3			Sampling (Issues, Methods, Effects)	10	Sampling		9, 10
13	Tue Oct 7	4.4 to 4.6			The Central Limit Theorem	11-EC	CLT/World (EC)	10, 11*	
14	Thu Oct 9	5.3		Estimation	Confidence Intervals	12	CI for Intervals		11*, 12
15	Tue Oct 14		HW3		CIs for Proportions	13	CI for Proportions	12, 13	
16	Thu Oct 16	6.1 & 6.4			Hypotesting & Zs	14	Writing Hypotheses		13, 14
17	Tue Oct 21				Hypotesting for Large ns	15	Two Tests	14, 15	
18	Thu Oct 23	6.3 & 6.8	HW4		The "t" test, for small ns	16	CI & Test Ages		15, 16
19	Tue Oct 28	5.4			Sample Size Estimation	17	Estimating n Needed	16, 17	
20	Thu Oct 30	7.1, 7.3, & 10.1		Covariation	Differences in Means	18	Comparing Means		17, 18
21	Tue Nov 4	7.2	HW5		Differences in Proportions	19	Comparing Proportions	18, 19	
22	Thu Nov 6	12.1			Analysis of Variance	20, 21-EC	ANOVA (+ MODELS EC)		19, 20, 2
	Tue Nov 11				(Veteran's Day)				
23	Thu Nov 13	9.4 & 9.5			Scatterplots & Correlation	22	Grade Correlations		22
24	Tue Nov 18	9.1 to 9.3	HW6		Regression	23	Regression Lab	20,21*,22,23	
25	Thu Nov 20	10.2 & 11.1			Multiple Regression	24-EC	Multiple Reg (EC)		23, 24*
26	Tue Nov 25	8.1	HW7	Association	Crosstabulations	25	TBA (any)	24*, 25	
	Thu Nov 27				(Thanksgiving)				
27	Tue Dec 2	8.2 & p.233			Dependence	26, 27-EC	TBA (SCU) (& 27-EC)	26, 27*	
28	Thu Dec 4	pp.238 to 243	HW8		Association	28-EC	Measures of Assoc (EC)		25,26,27*,2
x	Tue Dec 9				(no lecture - work session only)			28*	
x	Thu Dec 11				(no lecture - work session only)				
x	Thu Dec 18	-	HW9		(no meetings - deadline only - 10am, firm!)				

Probability & Z Scores

SOC424 – Statistics w/ Dr. Ellis Godard

Admin Inference Samples Curves Percentiles Z Scores Examples Scales

Fair warning…

LOTS of slides today…

*Like ,seriously… **LOTS!!!***

breathe deeply…

roll your head…

AAAAAaaaaahhhh…..

Ready? ☺

8 SOC424 w/ Dr. Ellis Godard -- Slide

Admin Inference Samples Curves Percentiles Z Scores Examples Scales

Announcements

- **Grading Updated (PDF & email)**
  - You've submitted ~17%, I've graded ~94% of that
  - *Still* need some intake forms & headshots
    - Some said "no" either to email or to pdf – email if you change your mind
- **Be cautious on homework!**
  - Read the questions, Don't skip questions, & Answer the Questions
  - Work alone! Don't "help" others by sharing your work
    - **Only** labs are group work – & not even all of those – TODAY's is **solo**!!
- **Lots of extra credit!** (Mystery Measurements etc.)
- **Coming Soon (not today): Best. Lab. Ever. Evah.**
  - All the pieces come together – light @ end of tunnel ☺

5 SOC424 w/ Dr. Ellis Godard -- Slide

Admin Inference Samples Curves Percentiles Z Scores Examples Scales

Review of the course so far…

- **Basic Terms**
  - Variables vs Values; Sample vs Population
- **Inference**
  - Want to generalize from a sample to a population
  - To do that, must first describe the sample
- **Measurements**
  - Descriptions depend on how we measured the sample
  - Nominal, Ordinal, and Interval descriptions differ
- **Sample Descriptions**
  - Central Tendency: Mean, Median, Mode
  - Dispersion: Std. Dev., Ranges, Variation Ratio
- **Distances from the Mean**
  - Sampling Error (unknown; don't know the parameters)
  - Standard deviations (z-scores; # of std devs from mean)
  - Outliers (e.g. more than 1.5xIQR from mean?)
- **Now: Distances are Associated w/ Probabilities**

9 SOC424 w/ Dr. Ellis Godard -- Z Scores

Outline for Today…

- Background:
  - **Inference**
  - Variation across **Samples**
  - **Curves** ("areas under the curve")
- Distances:
  - **Percentiles** (one more univariate distance)
  - **Z-Scores:** Standardized Differences
    - Meaning, calculation, and *lots* of
  - **Examples**
  - Standardized **Scales**
- Lab: Standardization & Z-Scores (2nd ***solo*** lab!)

From Description to Inference

Given information about a sample drawn from a population, what can we say about the characteristics of the population as a whole?

In other words, what is the relation between *sample statistics* and *population parameters?*

- Sample statistics are the *best estimate* of the corresponding population parameter.
- Likewise, we will use sample statistics to make *inferences* about population parameters.

Correspondence of Point Estimates

Sample Statistics	Population Parameters	
Mean ($\overline{Y}$)	Mean	$(\mu; \hat{Y})$
Mode	Mode	
Median	Median	
Range	Range	
Variance (s^2)	Variance	(σ^2)
Std. Deviation (s)	Std. Deviation	(σ)

Baseball Data (% of games won)

	Sample (n=10)	Population (N=26)
Mean	0.518	0.500
Median	0.522	0.487
Mode	0.500-0.549	0.450-0.499
Range	0.414-0.667	0.395-0.667
Variance	0.005	0.004
Std. Dev.	0.078	0.063

The Problem w/ Samples

Any sample, even if randomly drawn, may not be representative of the population.

A non-representative sample will lead us to make erroneous statements about the population.

With some rudimentary probability theory, we can determine how likely it is that we will draw a certain sample – and, given our sample, what the population probably looks like.

Probability Distributions

The *probability* of a particular outcome is the relative frequency that event can be expected to occur, the number of successful outcomes divided by the total attempts

It ranges from 0 to 1 (0% to 100%)

Note: "percent" is like two hidden decimals

The collection relative frequencies for each and every possible outcome is a *probability distribution*.

The probability distribution for a variable provides a listing of the probabilities of the various possible occurrences.

Probability of an Event

Can be obtained:

1. empirically
   - e.g. by actually flipping a coin.
2. theoretically
   - e.g. by assuming that a coin is "fair" -- that a head is as likely as a tail

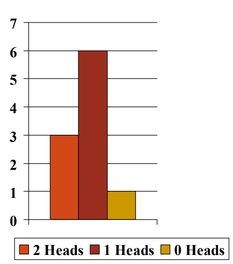
Example: Flipping Coins

Imagine the following experiment:

1. Toss two coins simultaneously and record the number of heads each time
   - 2 (head & head)
   - 1 (head and tail, or tail and head)
   - 0 (both tails)
2. Repeat step one 10 times
3. Create a relative frequency histogram of your results.

My Experiment's Outcomes (n=10)

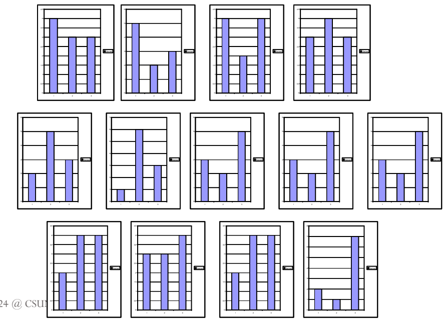
Value	freq
2H	3
1H	6
0H	1
Total	10



2H much more likely?

Are my coins "unfair"?

Coin Samples - Wide Variation



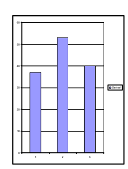
Law of Large Numbers

- As the number of times an "experiment" (e.g. flipping a coin) increases…

  …the ***empirical*** probability approaches the ***theoretical*** probability of an event.

Law of Large Numbers

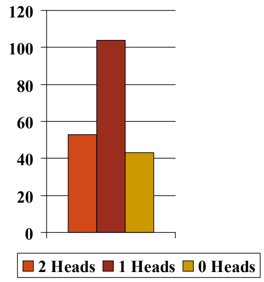
- Combining those into one large sample
- Distribution looks close to what a population of fair flips should look like:



- In general (and to an extent), larger samples better approximate population distributions

Larger Sample Outcomes (n=200)

Value	Freq.
2H	53
1H	104
0H	43
Total	200



Less difference here -

Did my coins get fair??

Flipping a Fair Coin

- If the coin is fair, the *theoretical probability* of a head is *exactly* 0.5.
- If the coin is fair, the *empirical probability* of a head *approximates* 0.5 as the number of experiments increase.
- Thus, *if* we assume the coin is fair, we can determine theoretically the likelihood of certain outcomes.

Expected Number of Heads When Two Coins are Tossed Simultaneously

Y	P(Y)
0H	0.25
1H	0.50
2H	0.25
	1.00

Note that the sum of the expected probabilities of each values *always* equals one (1.0, or 100%)

Number of Children in a Family

Y	P(Y)
0	0.49
1	0.21
2	0.15
3	0.08
4	0.04
5	0.02
6+	0.01
Total	1.00

Such a distribution is determined *empirically*. To obtain it you would need to conduct a *census* of all the families in a population.

If you randomly selected one family from this population what is the probability that it has 2 children? 4 *or more* children?

Continuous Random Variables

- So far, we have defined probability as:

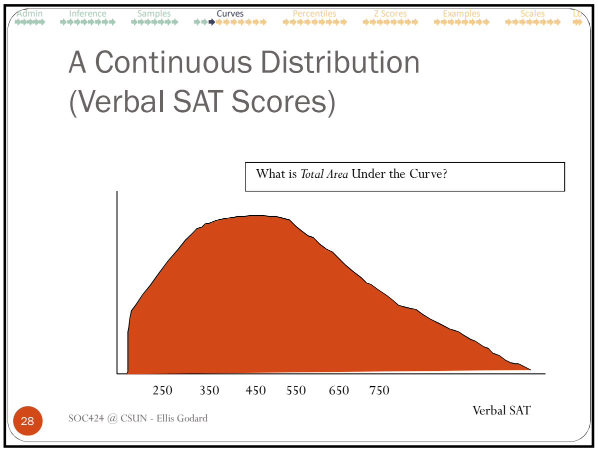
$$P = \frac{\text{number of successful outcomes}}{\text{total number of outcomes}}$$

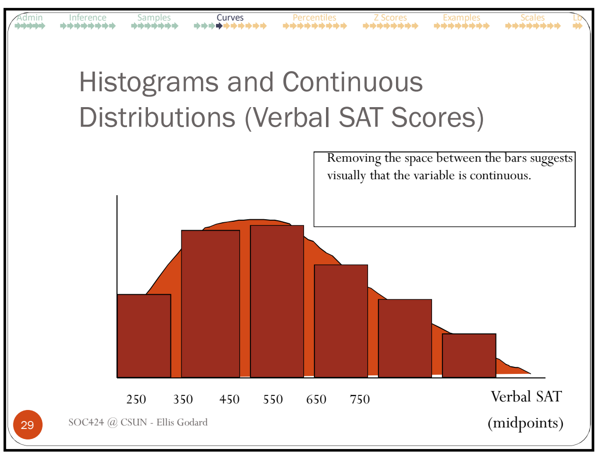
- As we move to variables that can take on an infinite set of values, this definition is changed to areas of the probability distribution. Now

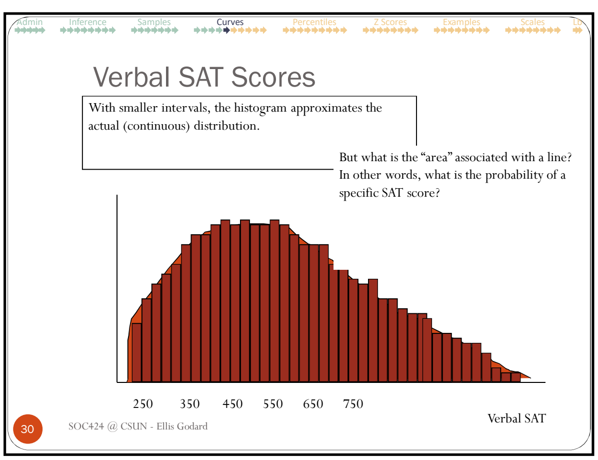
$$P = \frac{\text{area under certain portion of curve}}{\text{total area of the curve (which equals 1.0)}}$$

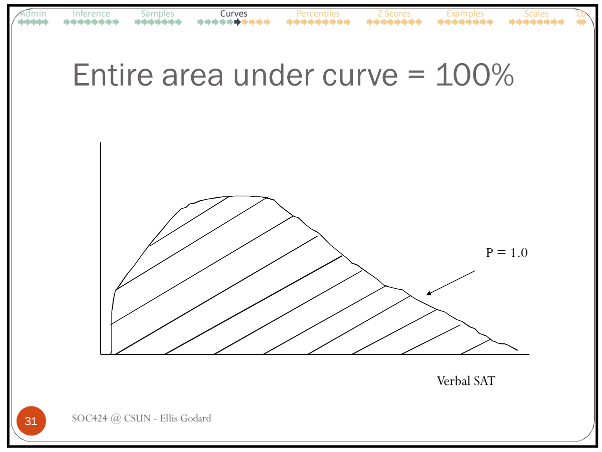
Continuous Probabilities

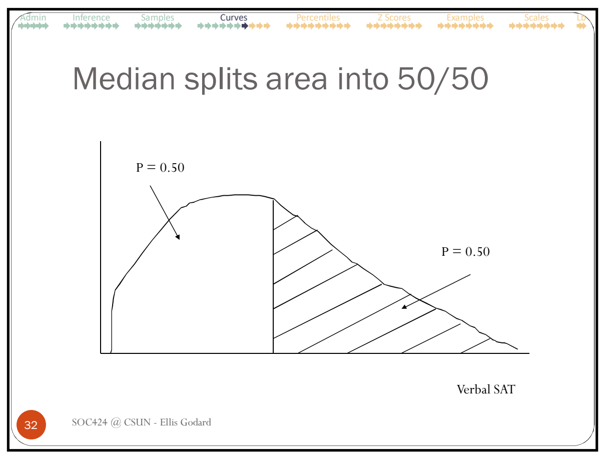
- For such distributions, probabilities can be assigned only to *intervals of numbers* not to a specific value.
- For example, we might be interested in the probability that family income is *above $50,000* (not that it *is* $50,000), or that verbal SAT score are *between 400 and 500* (but not a particular score).

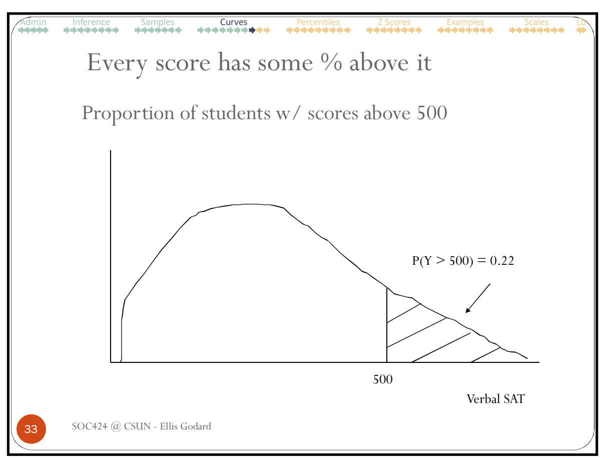

A Continuous Distribution
(Verbal SAT Scores)
What is *Total Area* Under the Curve?
250 350 450 550 650 750
Verbal SAT
28
SOC424 @ CSUN - Ellis Godard


Histograms and Continuous
Distributions (Verbal SAT Scores)
Removing the space between the bars suggests visually that the variable is continuous.
250 350 450 550 650 750
Verbal SAT
(midpoints)
29
SOC424 @ CSUN - Ellis Godard


Verbal SAT Scores
With smaller intervals, the histogram approximates the actual (continuous) distribution.
But what is the "area" associated with a line? In other words, what is the probability of a specific SAT score?
250 350 450 550 650 750
Verbal SAT
30
SOC424 @ CSUN - Ellis Godard

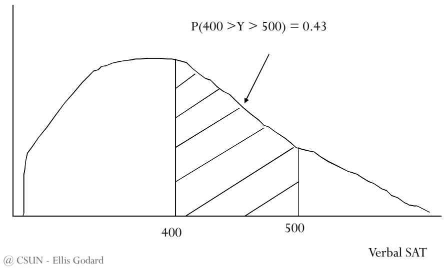

Entire area under curve = 100%
P = 1.0
Verbal SAT
31
SOC424 @ CSUN - Ellis Godard


Median splits area into 50/50
P = 0.50
P = 0.50
Verbal SAT
32
SOC424 @ CSUN - Ellis Godard


Every score has some % above it
Proportion of students w/ scores above 500
P(Y > 500) = 0.22
500
Verbal SAT
33
SOC424 @ CSUN - Ellis Godard

Some % is between every two

Proportion w/ scores between 400 and 500



Distances from the Mean

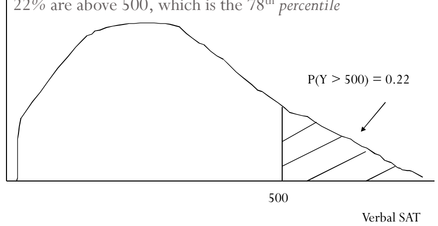
- **Previously: Sample Differences**
  - Sampling error from population mean
  - Range +/- one std. dev.
    - 68% of cases, if a normal curve
  - Outliers are 1.5xIQR from middle (by *one* definition)
- **Next: Standardized differences**
  - Individual scores ("z score")
  - Associated Probabilities (next time)
  - Area under the curve (e.g. 68% w/i 1 stdev)
- First: Percentiles vs. Percentages

Percentiles vs. Percentages

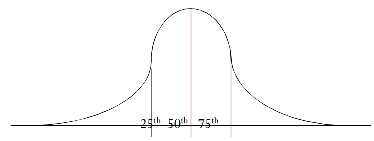
Percentile = location of a score

Percentage = area under the curve

22% are above 500, which is the 78th *percentile*



Percentiles you've already seen



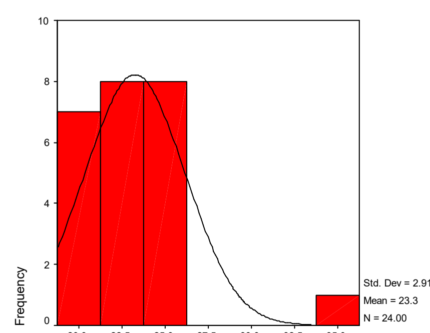
- 50th = the Median
- 25th to 75th = Inner-Quartile Range

Example: Class ages (frequencies)

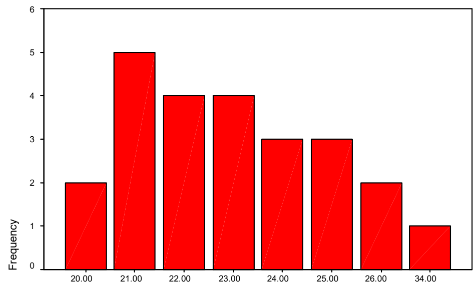
AGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20.00	2	8.3	8.3	8.3
	21.00	5	20.8	20.8	29.2
	22.00	4	16.7	16.7	45.8
	23.00	4	16.7	16.7	62.5
	24.00	3	12.5	12.5	75.0
	25.00	3	12.5	12.5	87.5
	26.00	2	8.3	8.3	95.8
	34.00	1	4.2	4.2	100.0
	Total	24	100.0	100.0	

Example: Class ages (*histogram*)



Example: Class ages (*bar chart*)



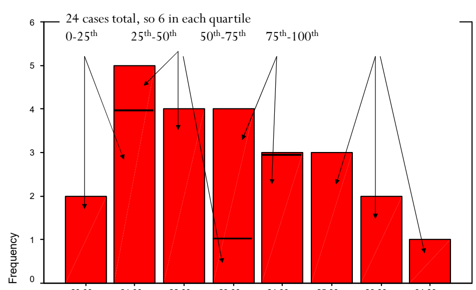
Example: Class ages (univ. stats)

Statistics

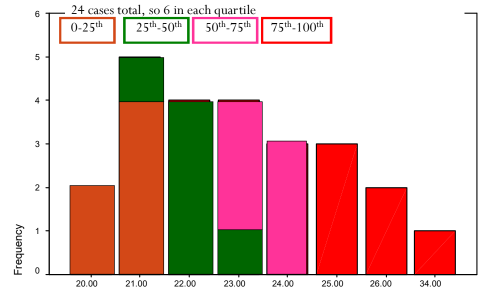
AGE

N	Valid	24
	Missing	0
Mean		23.2500
Median		23.0000
Mode		21.00
Std. Deviation		2.9080
Range		14.00
Percentiles	25	21.0000
	50	23.0000
	75	24.7500

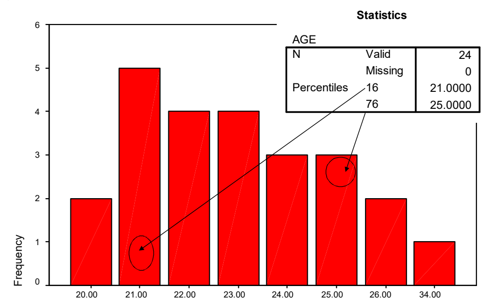
Example: Class ages (*percentiles*)



Example: Class ages (*percentiles*)



Any Score as a Percentile

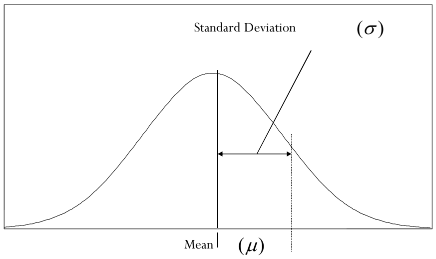


A New Kind of Distance

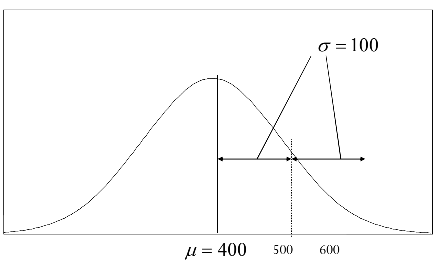
- **Percentiles measure distances in terms of the percent of scores between them**
  - Approximately 60% are between 21 and 25 years of age (the 76th & 16th percentiles)
- **Could instead measure the difference in terms of how many standard deviations they are away from each other**
  - There's a difference of four years from 21 to 25
  - The standard deviation is about 2.9
  - Four divided by 2.9 is about 1.38
  - 21 & 25 are about 1.38 standard deviations apart

STANDARD DEVIATION

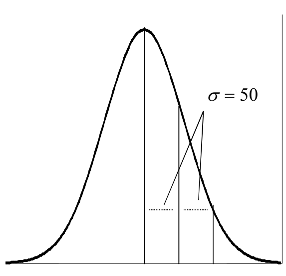
= "typical" deviation from the mean



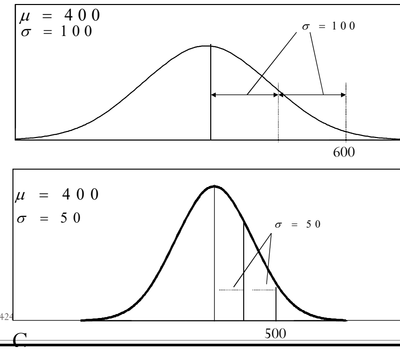
EXAMPLE: Verbal SAT



EXAMPLE: Math SAT



Both are *standard normal distributions*,
but one is much wider than the other.



• C

Z-SCORES: In theory

- Distance from the mean, measured as a number of standard deviations
- The z-score corresponding to any value (such as Y_i) is the number of standard deviations that that score (Y_i) is from the *population* mean ("Mu"):

$$z = \frac{Y_i - \mu}{\sigma}$$

Z-SCORES: In practice

- Two problems:
  - We rarely *know* the pop. mean
  - Data distributions often *not* normal
- Solution:
  - Calculate a z-score as an *estimated number of standard deviations from the mean*
  - Use the sample mean as a point estimate of the population mean:

$$z = \frac{Y_i - \bar{Y}}{\sigma}$$

Z-SCORES: In problems

- Can be used to talk about the position or location of any value, in relation to the curve or distribution of scores:

A specific value

Population Mean

$$z = \frac{Y - \mu}{\sigma}$$

Population Std. Deviation

EXAMPLE 1:

If the population mean is 400 and the population standard deviation equals 100 (in other words, $\mu = 400$ *and* $\sigma = 100$), then…

What is the z-score associated with Y_i=500?

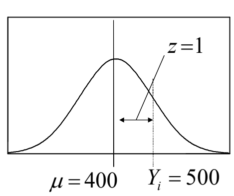
$$z = \frac{Y_i - \mu}{\sigma} = \frac{500 - 400}{100} = 1$$

Note: a negative z means that the value falls below the mean (150 < 400).

EXAMPLE 1:

$\mu = 400 \quad \sigma = 100 \quad Y_i = 500$

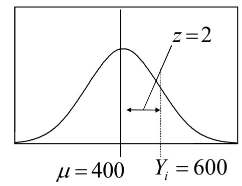
And here's the picture you should have drawn to try that (looking for the z-score associated w/ Y_i=500):



EXAMPLE 2:

$\mu = 400 \quad \sigma = 100 \quad Y_i = 600$

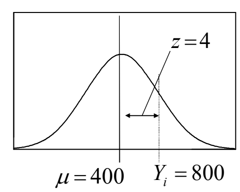
$$z = \frac{Y_i - \mu}{\sigma} = \frac{600 - 400}{100} = 2$$



EXAMPLE 3:

$\mu = 400 \quad \sigma = 100 \quad Y_i = 800$

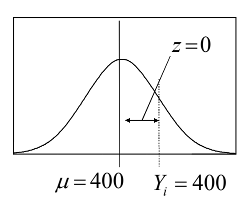
$$z = \frac{Y_i - \mu}{\sigma} = \frac{800 - 400}{100} = 4$$



EXAMPLE 4:

$\mu = 400 \quad \sigma = 100 \quad Y_i = 400$

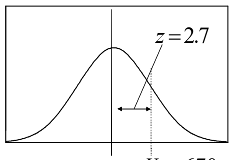
$$z = \frac{Y_i - \mu}{\sigma} = \frac{400 - 400}{100} = 0$$



EXAMPLE 5:

$\mu = 400 \qquad \sigma = 100 \qquad Y_i = 670$

$$z = \frac{Y_i - \mu}{\sigma} = \frac{670 - 400}{100} = 2.7$$

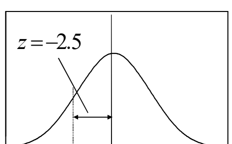


$\mu = 400 \qquad Y_i = 670$

EXAMPLE 6:

$\mu = 400 \qquad \sigma = 100 \qquad Y_i = 150$

$$z = \frac{Y_i - \mu}{\sigma} = \frac{150 - 400}{100} = -2.5$$



$Y_i = 150 \qquad \mu = 400$

Z-Scores for External Values

- **Another sample's value**
  - How different from Family A is a family without a television?
    - Z_0 = (0 – 10) / 6 = (-10)/6 = 1.66
  - The central tendency of Family A is 1.66 standard deviations from a family with no television viewing.
- **An imaginary value**
  - What would be the z-score for someone who watched 22 hrs/week?
    - Z_{22} = (22 – 10) / 6 = (12)/6 =2
  - They would be 2 standard deviations from this family's mean
- **An established standard**
  - If research calls more than 8 hrs unhealthy, how is Family A doing?
    - Z_8 = (8 – 10) / 6 = (-2)/6 = 0.33
  - This family exceeds the guidelines by 0.33 standard deviations

Standardizing a Scale

- Could convert *all* scores into z-scores
- Called a "standardized distribution"
  - Has a mean of 0
- If the original shape was normal, this new set of z-scores is a "standard normal distribution"
  - Has a standard deviation of 1

Standardization vs. Normalization

- Standardization means you convert scores from one scale of measurement into another using a consistent formula
  - Such as by subtracting the mean of the distribution from every score and dividing by the distribution's standard deviation).

$$z_i = \frac{X_i - \mu}{\sigma}$$

- Normalization means that force the scores into a bell-curve, regardless of whether or not that transformation is consistent
  - It ignores the actual shape of the distribution of scores, pretending there are no outliers, no skew, etc.

Z-SCORES STANDARDIZE

- Standardized in terms of distance from the mean
  - Z score = $(Y_i - \bar{Y}) / s$
  - = # of standard deviations a score is from the mean
- For example, Family A (from previous lecture)
  - Y_i={0, 4, 8, 12, 16, 20} ; $\bar{Y}$=10; s =6

Z_0: (0-10)/6 = -10/6 = -1.66	**Z_{12}: (12-10)/6= 2/6 = 0.33**
Z_4: (4-10)/6 = -6/6 = -1	**Z_{16}: (16-10)/6 = 6/6 = 1**
Z_8: (8-10)/6 = -2/6 = -0.33	**Z_{20}: (20-10)/6 = 10/6 = 1.66**

- So, the standardized (z) scores for Family A are:
  - -1.66, -1, -0.33, 0.33, 1, and 1.66

PROPERTIES OF Z's

- Note that z scores sum to zero:
  - -1.66 + -1 + -0.33 + 0.33 + 1 + 1.66 = 0
- More importantly, any z-score distribution (large sample, normal distribution)…
  - …has a *mean* of zero (here, 0/6=0)
  - …is normal (symetric & bell-shaped)
  - …has a standard deviation of 1 (here, ~1.23)

Example: Pts missed on HW1

HW1

		Frequency	Percent
Valid	1.50	1	4.3
	1.75	1	4.3
	2.25	1	4.3
	2.75	1	4.3
	3.50	5	21.7
	4.25	2	8.7
	4.50	1	4.3
	5.00	1	4.3
	5.25	1	4.3
	5.75	1	4.3
	6.00	1	4.3
	6.25	1	4.3
	6.50	1	4.3
	6.75	1	4.3
	7.75	1	4.3
	8.00	1	4.3
	8.75	1	4.3
	10.00	1	4.3
	Total	23	100.0

- Basic statistics:
  - Mean = 4.9891
  - Stdev = 2.2582
- 5 students missed 3.5 pts
  - = .6594 stdevs from mean

$$\frac{4.9891 - 3.5}{2.2582} = 0.6594$$

  - Z-score for 3.5 = 0.6594

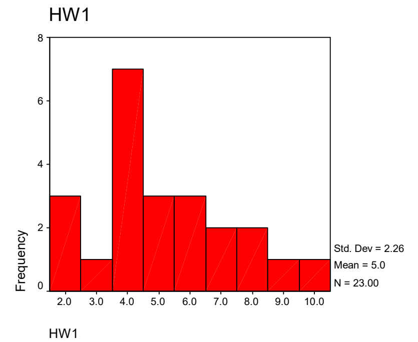
Example: Pts missed on HW1

Zscore(HW1)

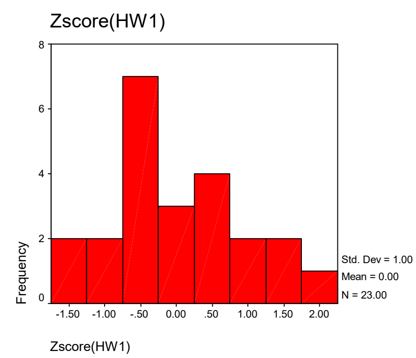
		Frequency	Percent
Valid	-1.54512	1	4.3
	-1.43441	1	4.3
	-1.21299	1	4.3
	-.99157	1	4.3
	-.65944	5	21.7
	-.32731	2	8.7
	-.21661	1	4.3
	.00481	1	4.3
	.11552	1	4.3
	.33694	1	4.3
	.44765	1	4.3
	.55836	1	4.3
	.66907	1	4.3
	.77978	1	4.3
	1.22262	1	4.3
	1.33333	1	4.3
	1.66545	1	4.3
	2.21900	1	4.3
	Total	23	100.0

- This is a standardized distribution of the pts missed
- All of the scores have been converted to z-scores
- Your grades are computed in a similar fashion

Comparative Distributions



Comparative Distributions



Standardizing a Scale in SPSS

- ANALYZE > DESCRIPTIVES > DESCRIPTIVES
  - Note this SOLE exception. Usually >FREQUENCIES!
- Choose variable (from left list, to box on right)
- Click "save standardized values as variables"
- SPSS creates
  - New variable (e.g. "zage" for "age")
  - New column of data (z scores for each case)

Next Lab: Standardize Scores

SOLO!

<u>You should each do this separately. On your own. Not in groups. It's a "solo" lab.</u>

1. Choose an interval variable from the given dataset
2. Calculate *<u>your</u>* z-score for that variable (the z-score for *your value*)
3. Standardize the entire scale (*<u>not</u> by hand*! see previous slide!)
4. Report mean & standard deviation for that standard scale
   (again, *not* by hand – use SPSS, though you wont need it)
   (and *not* for the original variable – e.g. zage, not age!)
5. You don't need to submit any SPSS output
   - Just submit answers to 1, 2, & 4
   - And you should know the answers for 4 before you even start ☺