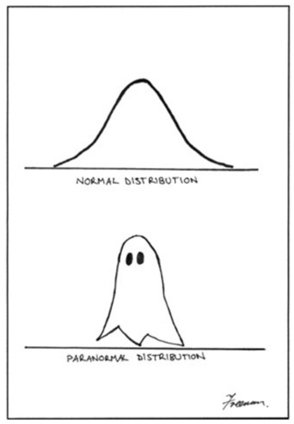
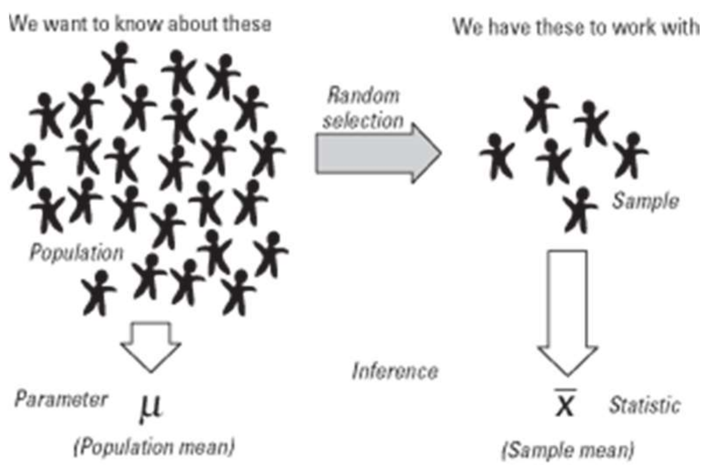
Sampling Distributions, Central Limit Theorem and the Normal Curve



Population vs. Sample



Parameter vs. Statistic

Statistic vs Parameter

	Mean Symbol	Mean Calculation	Standard deviation Symbol	Standard Deviation Calculation
Population	μ	$\frac{\sum_{i=1}^{n} x_i}{N}$	σ	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$
Sample	$\bar{X}$	$\frac{\sum_{i=1}^{n} x_i}{n}$	s or s_x	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$

- (n-1) because we already used up one “information point” (degree of freedom) by estimating the sample mean
- The larger the sample, the less impact that “-1” has

Sample Representativeness

- Being able to make conclusions about population from the sample → generalizability
- Need a representative sample: a sample that “looks like” population
- Best samples = probability samples: the likelihood of selection is known for every population element

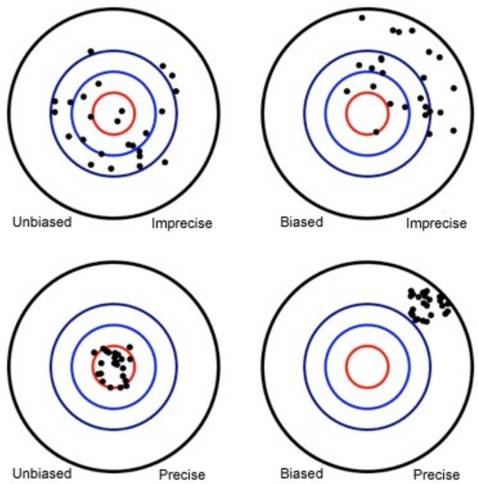
Statisticians fall asleep faster by taking a random sample of sheep.

Sampling Error

- Measure of representativeness
- Sampling error = difference between the sample statistic and population parameter

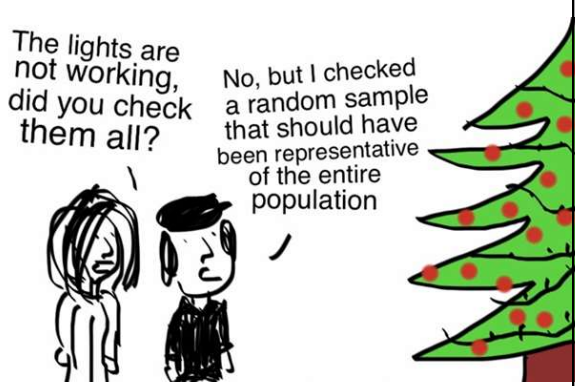
Components of Sampling Error

- Sampling error = random sampling error + systematic sampling error (bias)



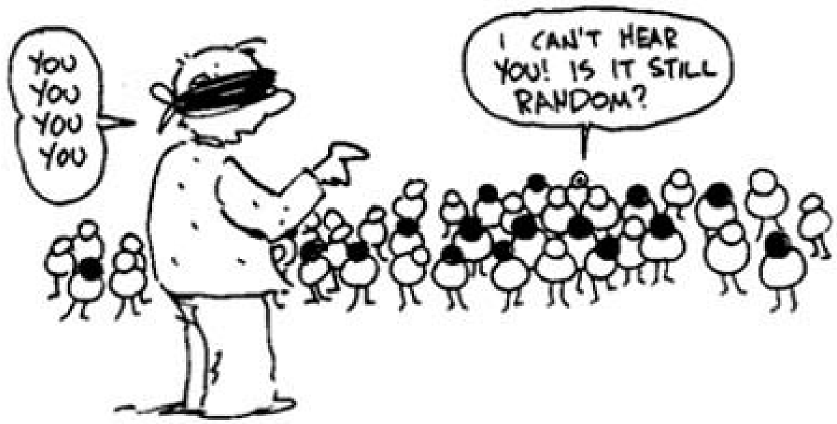
Probability Samples and Error

- Probability samples still have sampling error → random sampling error (due to chance)
- They should not have systematic sampling error (bias)



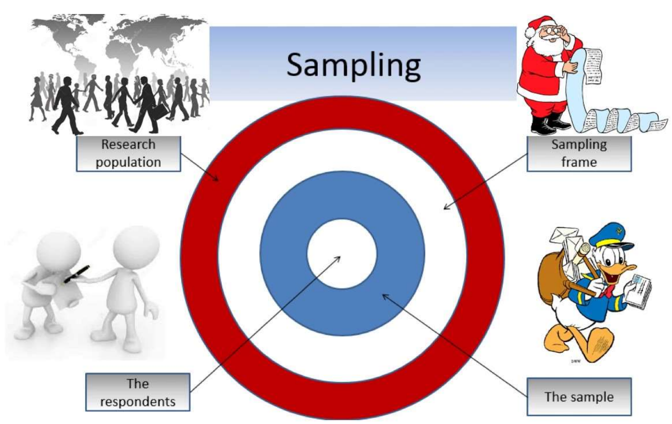
Three Main Reasons for Systematic Sampling Error (Bias)

1. The sample is non-random



Three Main Reasons for Systematic Sampling Error (Bias)

2. The sampling frame does not include all population



Three Main Reasons for Systematic Sampling Error (Bias)

3. There is non-response bias

©Alex Hallatt

Main Sources of Random Sampling Error: 1. Sample Size

The sample is small → random sampling error is large

- Error decreases with sample size
- That means absolute sample size, <u>not</u> sample size proportionate to size of population
- E.g., 1000 students will represent 20,000 students equally well as 1000 people will represent U.S. population
- But careful with subgroups!

Main Sources of Random Sampling Error: 2. Population Diversity

- The population is diverse (no homogeneity) → random sampling error is large

Homogenous population:

Heterogenous (diverse) population:

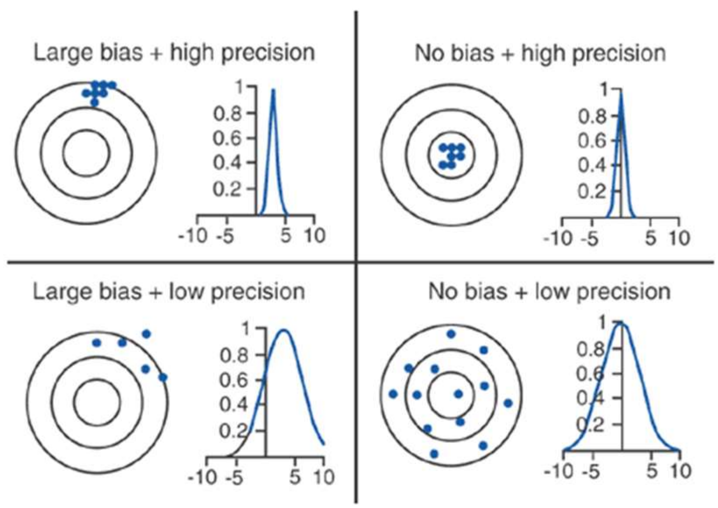
Focus on Random Sampling Error

- Aim to avoid both types of error when collecting data
- When analyzing, we assume there is no bias but address random sampling error (chance)
- To generalize from sample statistic to population parameter → need to take chance into account
- Still, we can never be 100% certain that the outcome is on target – only with a certain probability

Sampling Distribution

- One sample = one shot at a target that's our population parameter, e.g., mean
- Keep drawing many times → infinite number of samples (or all possible samples) → each time we have a mean
- Distribution of all such means = sampling distribution of the mean
- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Sampling Distribution



Central Limit Theorem

- If we know the sampling distribution → we know how much we can trust our estimate (from one sample)
- How can we get it? Cannot generate an infinite number of samples
- We use distribution generated by statistical theory → Central Limit Theorem
- http://vimeo.com/75089338
- So Central Limit Theorem tells us something about the (a) mean, (b) spread and (c) shape of a sampling distribution

Central Limit Theorem: Mean

- In the long run, $\overline{X} = \mu$
- If the average of $\overline{X}$ in all possible samples is μ, then $\overline{X}$ is an unbiased estimator
- Unbiased = the estimator on average is equal to the population mean
- If $\mu = 50$, but our $\overline{X} = 30$ → is this a biased estimate of the population mean?
- No! A single sample cannot prove that → we establish this biasness in the "long run"
- As long as on average, this estimation is correct, a single estimate can be any number (but with different probability)

Central Limit Theorem: Spread

- Standard error of the mean (SEM)
- This is standard error of the mean, not of the variable itself!

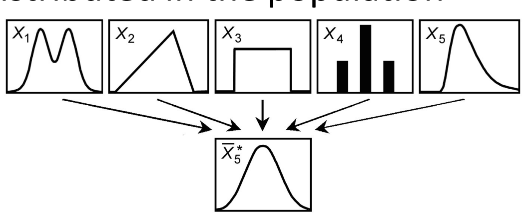
$$\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}$$

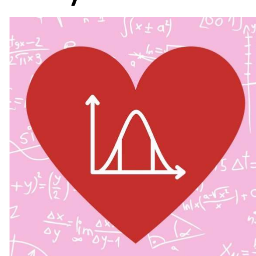
$\sigma_{\bar{X}}$ = standard error of the mean
s = standard deviation of the variable
n = sample size

- What makes this standard error smaller (and our estimate more precise)?
  - Larger sample size (if we want to reduce SEM by half, need sample size 4 times larger)
  - More homogenous population

Central Limit Theorem: Shape

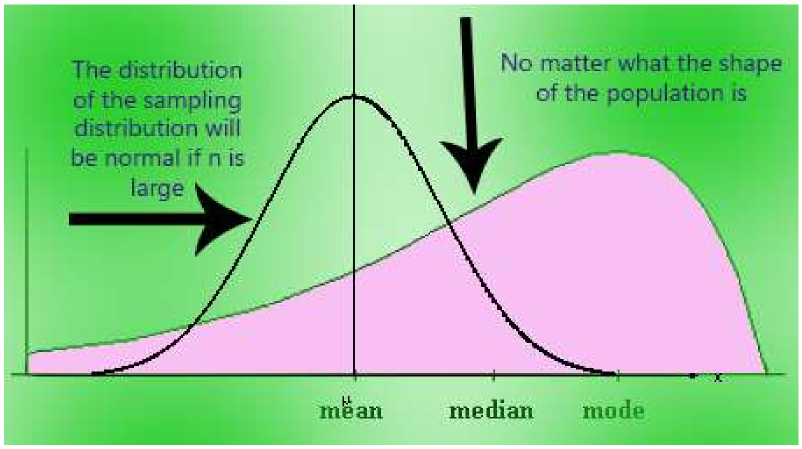
- Repeatedly draw random samples & get mean → distribution of the mean will be normal
- True even if the variable is not normally distributed in the population

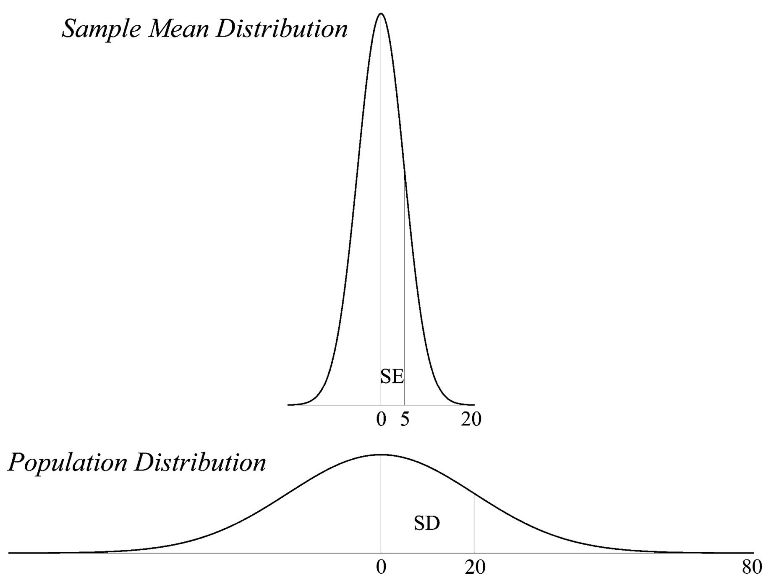




- http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- http://students.brown.edu/seeing-theory/probability-distributions/index.html

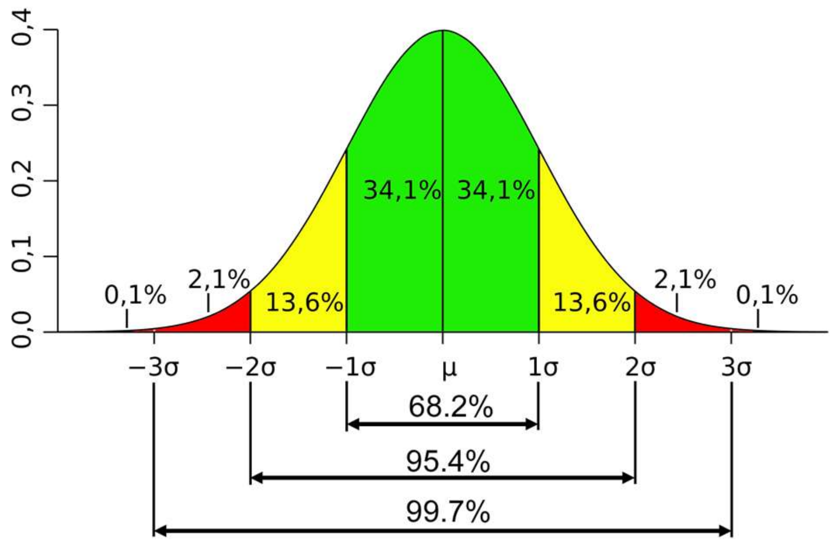
Central Limit Theorem (CLT)



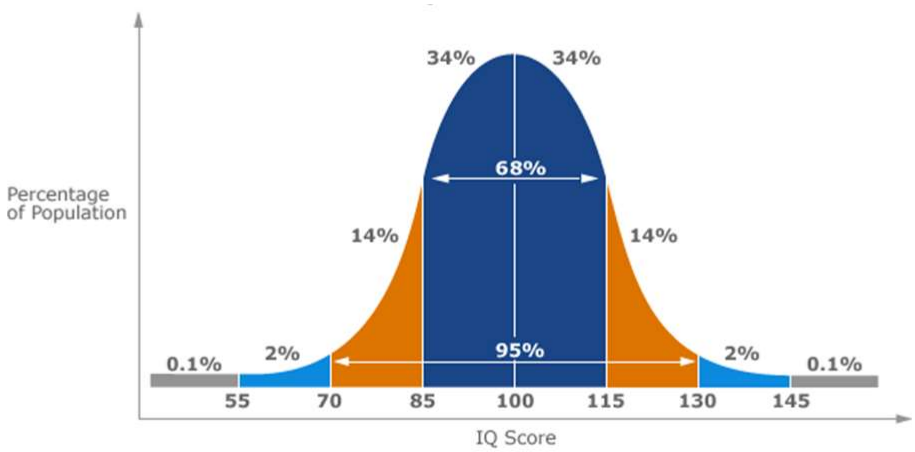


Sample size = 16, SEM = S/sqrt(n)= 20/sqrt(16) = 5

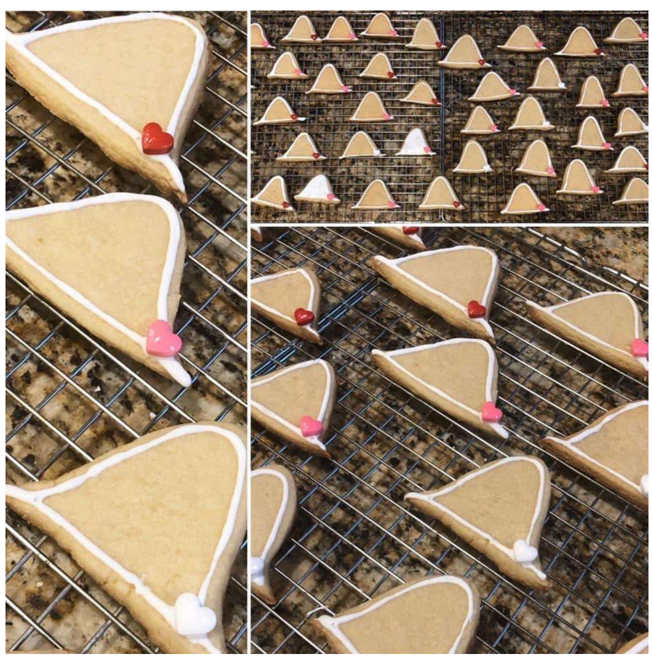
Properties of a Normal Distribution

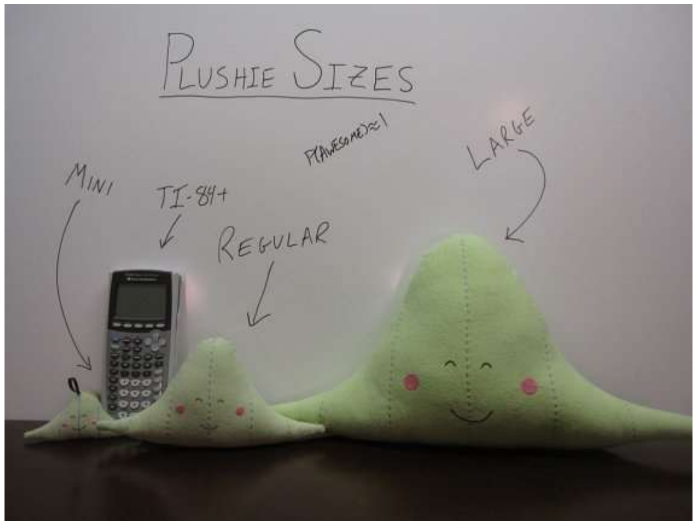


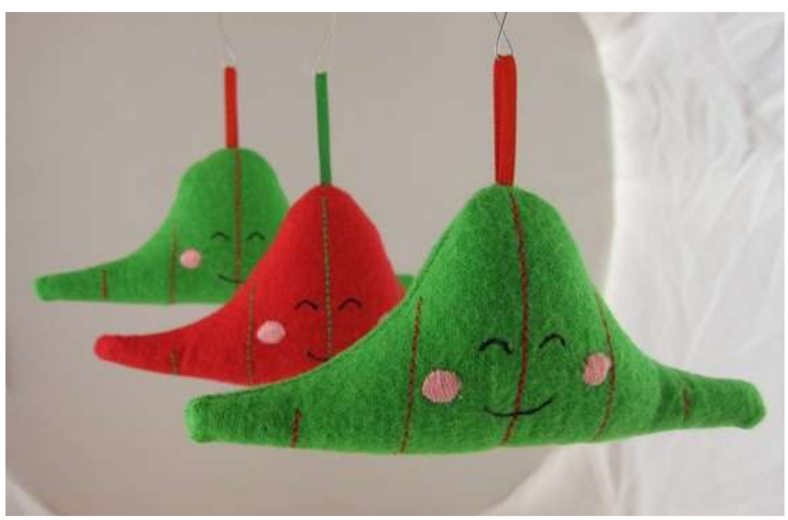
IQ Score Distribution



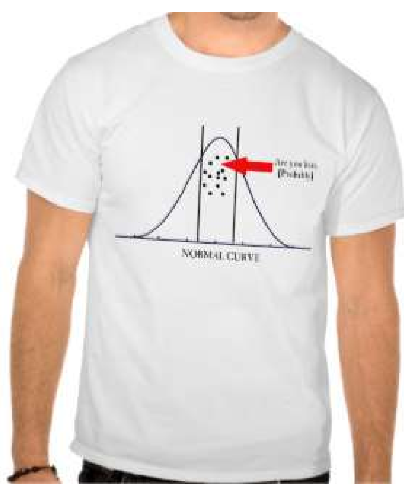
- Mean = 100, SD=15

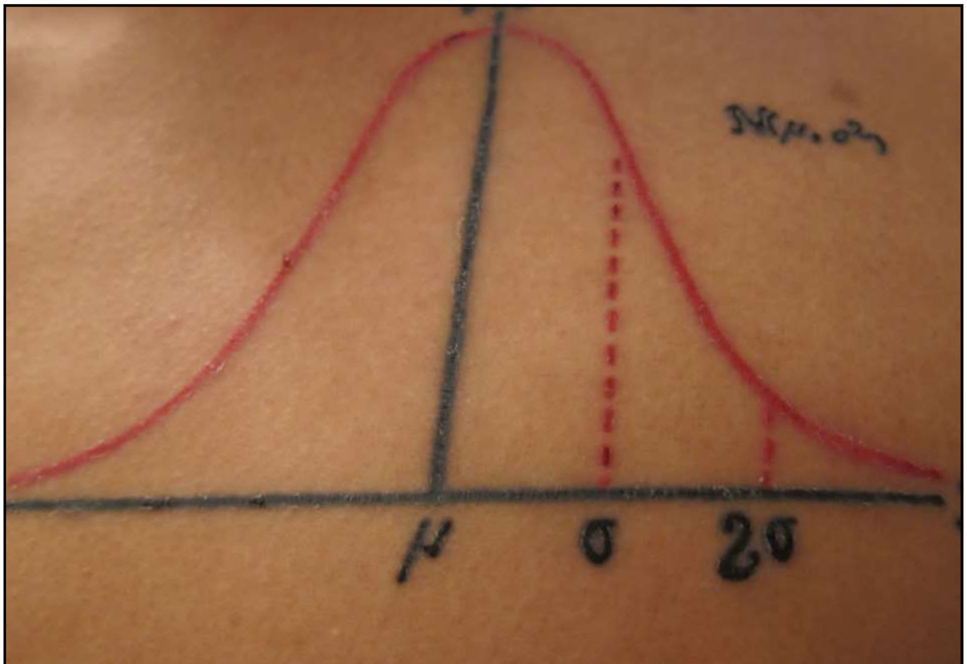



PLUSHIE SIZES
MINI
TI-84+
REGULAR
LARGE



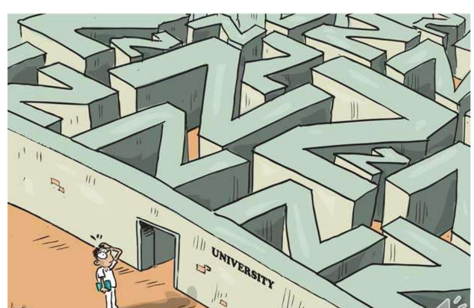

Statistically,
I'm Cute.

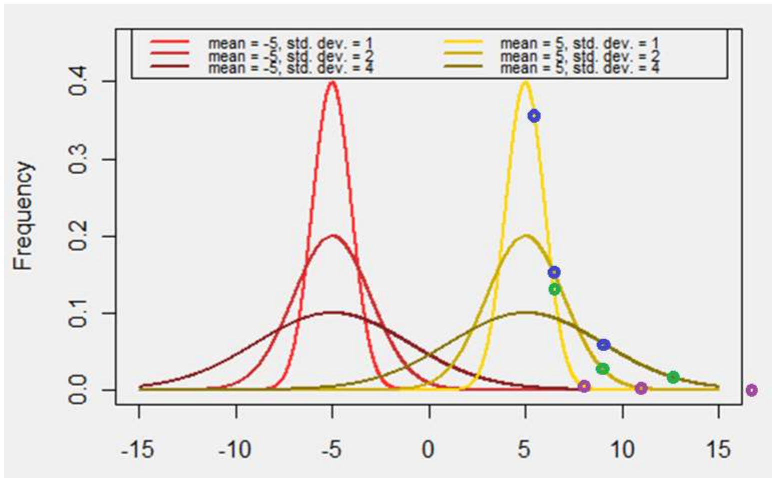

NORMAL CURVE



Z-Scores

- To compare data from normal distributions with different means and SD → use standardized scores
- Z-scores – measured in standard deviations
- 1 = 1 SD above mean
- -1 = 1 SD below mean





blue: z=1, green: z=2. purple: z=3

Calculating Z-Scores

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

- z = z-score
- X = individual score
- $\overline{\text{X}}$ = mean
- S = standard deviation

Example

- Given a normal curve with mean of 30 and a standard deviation of 2, find the z-score for X=25

z = (25-30)/2 = -2.5

- Given that X is normally distributed with mean of 20 and a standard deviation of 4, find the value of X corresponding to z-score = -2.5

$$X = \bar{X} + z*s$$

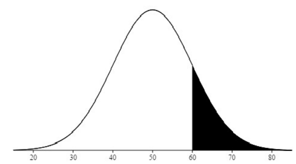
X = 20 - 4*2.5 = 10

Table B1 (in Salkind): Areas Beneath the Normal Curve

- Shows what percentage between the mean (z=0) and a given z-score
- However, you will often need to find the percentage above your value of z, below your value of z, or between two values of z
- To calculate, remember:
  - draw a picture to not get confused
  - % of scores that fall under the entire curve is 100%
  - % of scores that fall under any half of the curve is 50%

How Unusual is Our Observation?

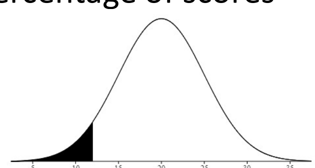
- Problem: Given a normal curve with mean of 50 and a standard deviation of 10, find the percentage of scores that will fall above a score of X=60.



- *Step 1*: Convert X to z: $z=(X-\bar{X})/s= (60-50)/10 = +1$
- *Step 2*: Enter Table B1 with z= +1.00 and read 34.13.
- *Step 3*: Since the percentage between z=0 and z= +1.00 is 34.13, the proportion above Z= +1.00 must be:
- 50.00 – 34.13 = 15.87% fall above 60

Another Example: Negative Side

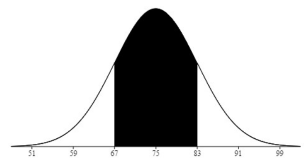
- Problem: Given a normal curve with a mean of 20 and a standard deviation of 5, find the percentage of scores below X=12.



- *Step 1*: Convert X to z: $z=(X-\bar{X})/s = (12-20)/5 = -1.6$
- *Step 2*: Enter Table B1 with z = -1.6. Only the positive half of the curve in the table → we enter with +1.6 → percentage between z=0 and z= +1.6 is 44.52
- *Step 3*: The amount above z=+1.6 is 50.00 – 44.52 = 5.48
- Because the normal curve is symmetric, the same percentage of cases will be above +1.6 as will be below -1.6, so the answer is 5.48% of the scores below X=12

Example with a Range

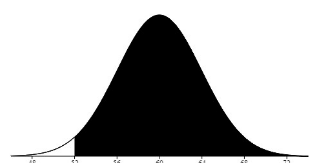
- Problem: Given a normal curve with a mean of 75 and a standard deviation of 8, find the percentage of scores between 67 and 83.



- *Step 1*: Convert both X=67 and X=83 to z scores. z=(X-$\bar{X}$)/s = (67-75)/8 = -1. And, z=(83-75)/8 = +1
- *Step 2*: For z = +1.00, the percentage between z=0 and z=1 is read to be 34.13 in the Table B1
- *Step 3*: Since the curve is symmetric, the same percentage will be found between z=0 and z= -1.00
- So the answer is 34.13 + 34.13 = 68.26% of the scores will be between 67 and 83

Starting from Below

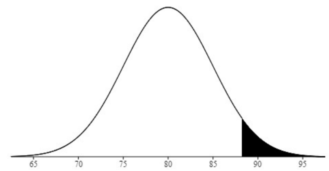
- Problem: Given a normal curve with a mean of 60 and a standard deviation of 4, find the percentage of scores above 52.



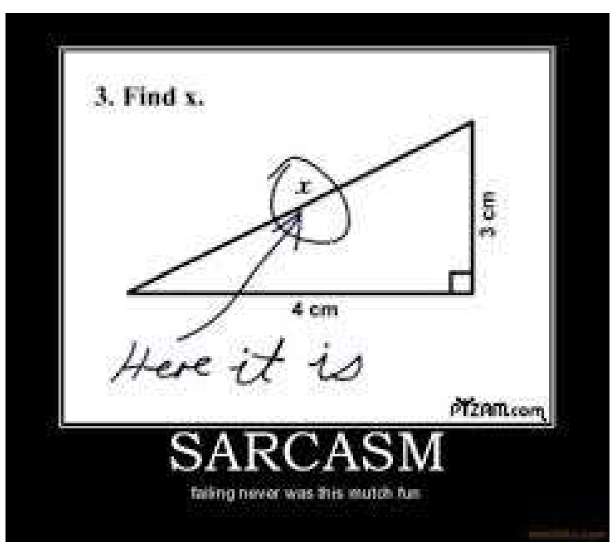
- *Step 1*: Convert X=52 to a z-score. z=(X-$\bar{X}$)/s = (52-60)/4 = -2. Note that this is below the mean, so we need both the area from z=-2 to the mean and the entire 50% above the mean.
- *Step 2*: The percentage between z=0 and z=2 is read to be 47.72 in the Table B1 . Since the curve is symmetric, the same percentage will be found between z=0 and z= -2.
- *Step 3*: So the answer is 50.00 + 47.72 = 97.72% of the scores will be above 52.

Now in Reverse: Find X

- <u>Problem</u>: Given a normal curve with a mean of 80 and a standard deviation of 5, find the X value such that only 5% of the cases are above it.

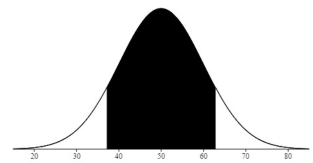


- *Step 1*: Top 5% → but Table B1 requires percentage of scores between the mean and the X → we enter with 50.00 – 05.00 = 45.00.
- Table B1: 45.00 → The closest entry is z = 1.65 or z=1.64; we can use either one, say, 1.65.
- *Step 2*: Convert z = 1.65 to X: X = $\overline{X}$ + z*s=80+(1.65*5) = 88.25.
- Answer: About 5% of the scores will exceed X=88.25.



Find X: Example with a Range

- Problem: Given a normal curve with a mean of 50 and a standard deviation of 10, find the 2 values of X that include the middle 80% of the distribution.



- *Step 1*: Middle 80% = 40% on the right and 40% on the left
- Enter the table with percentage 40.00. The closest to 40.00 is 39.97 for z=1.28. Only 10% will fall above that point.
- Because the curve is symmetric, only 10% will fall below the point z = -1.28. That means that the two points that include the middle 80% will be z = -1.28 and +1.28.
- *Step 2*: z values (-1.28 and +1.28)→X values. Using X= $\bar{X}$ + z*s: X_1=50+(-1.28*10)=37.2 and X_2=50+(1.28*10)=62.8. About 80% of the curve will fall between 37.2 and 62.8.