Statistical Models:
The Normal Distribution and Sampling Distributions

PSYC 203

Agenda

- Description to Inference
  - z-scores
  - Normal Distribution
- Choosing the Right Model
  - Normal or Nonnormal?
  - Linear or Nonlinear?
- Simple Modeling Concepts & Fit
  - Sampling Distributions
  - Confidence Intervals

Standard Scores (z-scores)

- Converts raw scores to a common metric

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

Characteristics of z-scores

- Magnitude & Sign
- Mean is always 0, standard deviation is always 1.

Computing z-scores

	Testing	History	Social
Student	**35**	**70**	**150**

Computing z-scores

	Testing	History	Social
Student	35	70	150
Test Mean	**25**	**80**	**150**

Computing z-scores

	Testing	History	Social
Student	35	70	150
Test Mean	**25**	**80**	**150**
Test SD	**5**	**10**	**40**

Computing z-scores

z score $\frac{35-25}{5} = 2$

	Testing	History	Social
Student	35	70	150
Test Mean	25	80	150
Test SD	5	10	40
z-score	**2.00**	**-1.00**	**0**

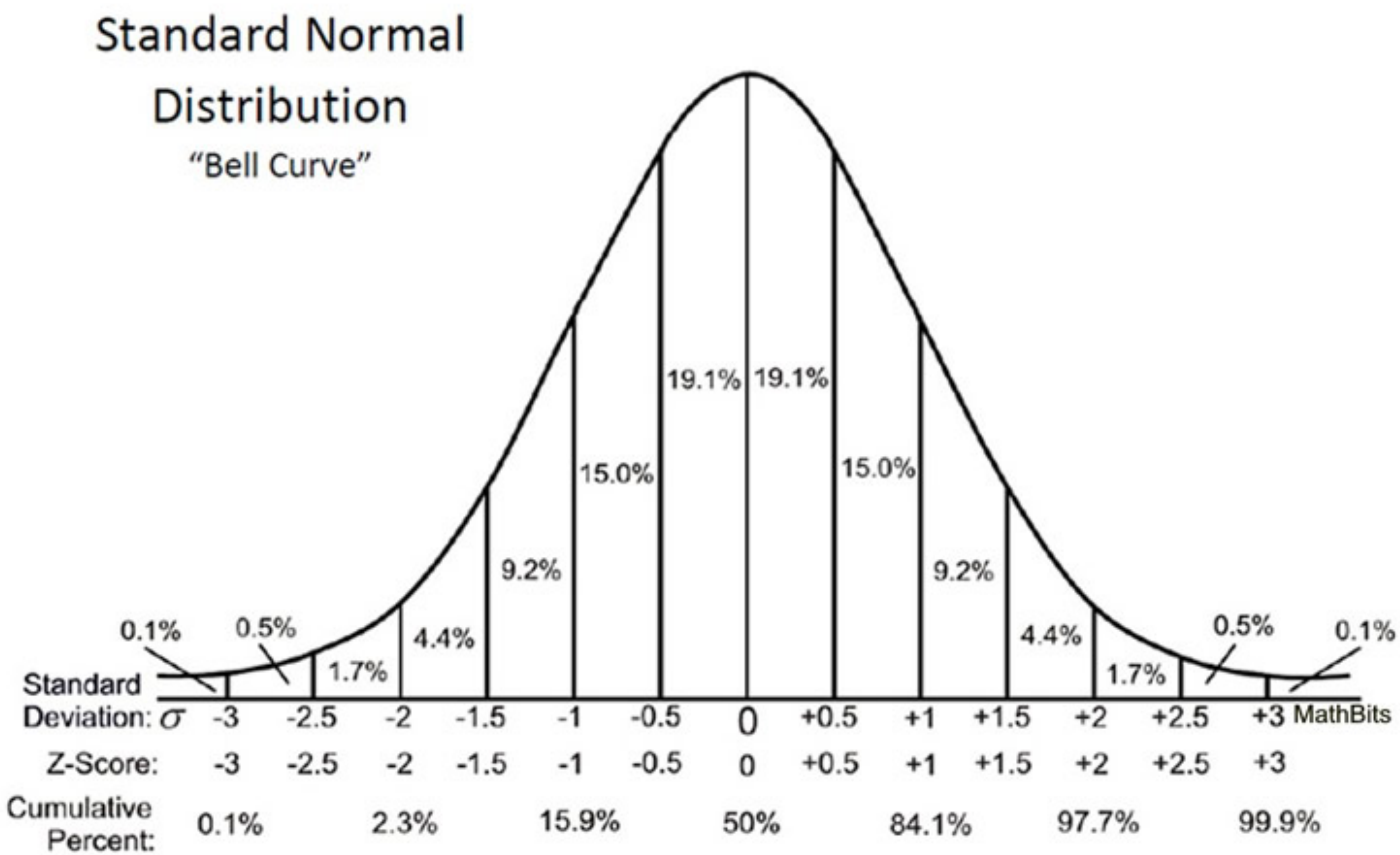
Using z-scores

- T-scores
  - T = 50 + 10(z)
  - Scale has a Mean of 50 and SD of 10
- IQ
  - IQ = 100 + 15(z)
  - Scale has a Mean of 100 and SD of 15
- PSYC 203 scores
  - PSYC = 200,000 + 10,000(z)
  - Scale has a Mean of 200,000 and SD of 10,000

Frequency Distributions and the Normal Curve

- Observed frequency distributions/histograms tell us what the observed data ‘are’
  - descriptive
- The normal curve is a mathematical distribution (a probability density function) with precise properties that can express how data ‘should be’
  - probability and inference
- Observed data often approximate a normal distribution, but don’t frequently perfectly conform.
- Serves as a foundational model because many variables display relatively normal distributions when collected from large samples.

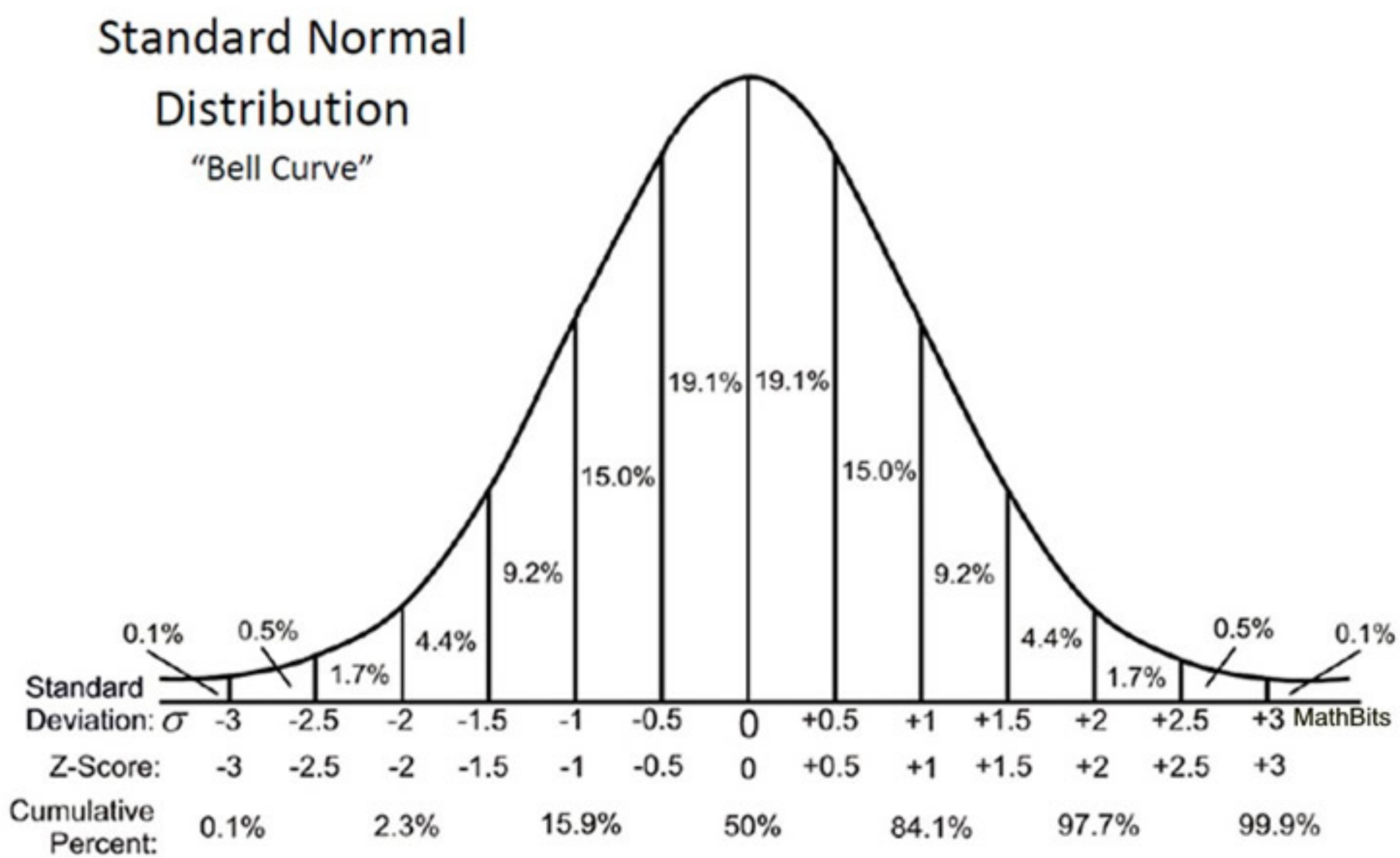
Normal Curve



Characteristics of the Normal Distribution

- Unimodal
- Symmetric
- Mode, Median, and Mean are Equal
- Asymptotic
- Two normal distributions DO NOT necessarily have same means and variances

Probabilities and Normality



http://davidmlane.com/hyperstat/normal_distribution.html

Standard Normal Distribution

- Can use this to answer
    - proportion of scores between Mean and a given score
    - proportion of scores above (below) a given raw score
    - proportion of scores between any two raw scores
    - what raw score falls above (below) a given proportion of cases

http://davidmlane.com/hyperstat/normal_distribution.html

Are my data normal?

- Visual inspection
  - Histogram
  - Q-Q plot
- Empirical tests

Using the Normal Distribution

- What if we knew the following:
  - The average test score on a given test for the population of high school seniors is 25.00 and the standard deviation is 5.00.
  - What is the probability of selecting a student who has a score of at least 32?

Using the Normal Distribution

- First, stated in terms of probability:
- p > 32 = ??
- Identify the exact position of X = 32 by converting to a z-score:

$$z = \frac{X - \bar{X}}{s} = \frac{32 - 25}{5} = \frac{7}{5} = 1.4$$

- What probability corresponds to z = 1.4?

http://davidmlane.com/hyperstat/normal_distribution.html

Another Example

- In my restaurant, I know that the average number of diners for my special Sunday brunch is 241 with a standard deviation of 19.

- One Sunday, I count 198 diners.

- If my distribution of Sunday diners is normal, what is the probability of having only 198 diners (or fewer)?

Agenda

- Description to Inference
    - z-scores
    - Normal Distribution
- **Choosing the Right Model**
    - **Normal or Nonnormal?**
    - **Linear or Nonlinear?**
- Simple Modeling Concepts & Fit
- Sampling Distributions
- Confidence Intervals

Agenda

- Description to Inference
    - z-scores
    - Normal Distribution
- Choosing the Right Model
    - Normal or Nonnormal?
    - Linear or Nonlinear?
- **Simple Modeling Concepts & Fit**
- Sampling Distributions
- Confidence Intervals

Models

- Everything we do in psychology can be expressed as:

$$obs_i = model + error_i$$

- What is the fit of our model under various scenarios?
- Does the model fit the data?
- Imagine we ask how many pets psychology faculty have?
    - One person: 2
    - Two people: 2, 2
    - Three people: 2, 2, 2
    - Four people: 2, 2, 2, 0

Model Fit

- Total Error (Sum of Squares)
    - Sum of the squared deviations (observed - model)
- Mean Squared Error
    - Divide SS by the degrees of freedom (n-1)

Agenda

- Description to Inference
  - z-scores
  - Normal Distribution
- Choosing the Right Model
  - Normal or Nonnormal?
  - Linear or Nonlinear?
- Simple Modeling Concepts & Fit
- **Sampling Distributions**
- Confidence Intervals

Sampling Distribution of Means

- We almost never know the 'true' distribution, so comparing observed data to an ideal normal distribution is often not reasonable.

- What we really want to know is,

    - "If there is a true population distribution, and we draw a given sample from that distribution, how much variability should we expect in samples drawn from the population?"

    - From this standpoint, we can determine the probability of any of our observed sample characteristics.

Sampling Distribution of Means

- As we have already seen, if we draw samples from a larger population, the means of those samples will form a normal distribution that clusters around the mean of the population.
    - **http://onlinestatbook.com/stat_sim/sampling_dist/index.html**
- The variability of the sampling distribution is less than the population variability.
- Sampling distribution variability is a function of sample size (larger sample = less variability).

Central Limit Theorem

- The central limit theorem states that given a distribution with a particular mean (μ) and variance (σ^2), the sampling distribution of the mean approaches a normal distribution with a mean (μ).

- The amazing and counter-intuitive thing about the central limit theorem is that no matter the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution.

Key Characteristics of Sampling Distribution of Means

- Normally distributed
- Mean = Population Mean
- Standard Deviation
- Standard Error of Mean $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{N}}$

How do we use this info?

- We know that the average score for adults age 25-55 on a national test of Batman trivia is 50 and the standard deviation is 15.
- If we were to draw 1000 samples of 25 individuals
- what would the mean of the 1000 samples be?
- what would be the standard error?
- What would happen if we drew the same number of samples, but each included 50 individuals?

Agenda

- Description to Inference
  - z-scores
  - Normal Distribution
- Choosing the Right Model
  - Normal or Nonnormal?
  - Linear or Nonlinear?
- Simple Modeling Concepts & Fit
- Sampling Distributions
- **Confidence Intervals**

Combining Concepts

- We have talked about estimating parameters
  - Compute the mean
- We have talked about the uncertainty associated with the parameter estimate
  - Sampling error
- We can put these two together
  - Confidence Intervals

Confidence Intervals

- How much uncertainty are we willing to tolerate?
  - 95%   Typical value chosen
  - 99%
- We know that 95% of cases in a normal distribution are between -1.96 and +1.96 z-scores
  - Great, but our data aren’t typically in standard scores, so we have to re-scale to the original metric

Computing Lower and Upper Bounds

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

$$-1.96 = \frac{X - \bar{X}}{s}$$

$$-1.96s = X - \bar{X}$$

$$\bar{X} + (-1.96s) = X$$

$$1.96 = \frac{X - \bar{X}}{s}$$

$$1.96s = X - \bar{X}$$

$$\bar{X} + 1.96s = X$$

Upper and Lower Bounds

- Our interval is based on sample means, NOT variation within a sample, so we will substitute the standard error for s

$$X_L = \bar{X} - 1.96SE$$

$$X_U = \bar{X} + 1.96SE$$

Two Important Points

1. We can use our knowledge of the normal curve to compute intervals of any size we like (68%, 95%, 99%).

2. We need to account for our relatively small samples, so we often multiple by something other than 1.96.

    - We will come back to this later in the course