STOR 455
STATISTICAL METHODS I

Jan Hannig

Fundamental Concepts (Section 1.2)

- Population: the entire group of individuals that we want information about.
- Sample: a part of the population that we actually examine in order to gather information.
- Sample size: number of observations/individuals in a sample.
- Statistical inference: to make an inference about a population based on the information contained in a sample.

Fundamental Concepts

- A *model* is mathematical description of the quantities of interest
  - Gaussian with unknown mean and variance
- A *parameter* is a value that describes the population. It's fixed but unknown in practice.
  - the mean and variance of the SAT score of all the students, who are about to take it.
- A *statistic* is a value that describes a sample. It's known once a sample is obtained.
  - The mean and variance SAT score of all the students, who are selected into the study.
  - A sample analogy of the parameter.
- Statistics is a course about lots of statistics!!! ☺

Models (Section 1.4)

- There are many possible model distributions
  - Gaussian distribution
  - Binomial distribution
  - Poisson distribution
  - Gamma distribution
  - …
- In this class we will almost exclusively use *Gaussian Distribution*

Density Curve

- Define a probability density function $f(x)$.
- The curve that plots $f(x)$ is called the corresponding <u>density curve</u>.
- $f(x)$ satisfies:
  - $f(x) >= 0$;
  - The total area under the curve representing $f(x)$ equals 1.
- Areas under the curve represent relative frequencies of observations

Normal Distribution (Section 1.4)

- Pictorially speaking, a Normal Distribution is a distribution that has a symmetric, unimodal and bell-shaped density curve.

- The mean and standard deviation completely specify the curve.

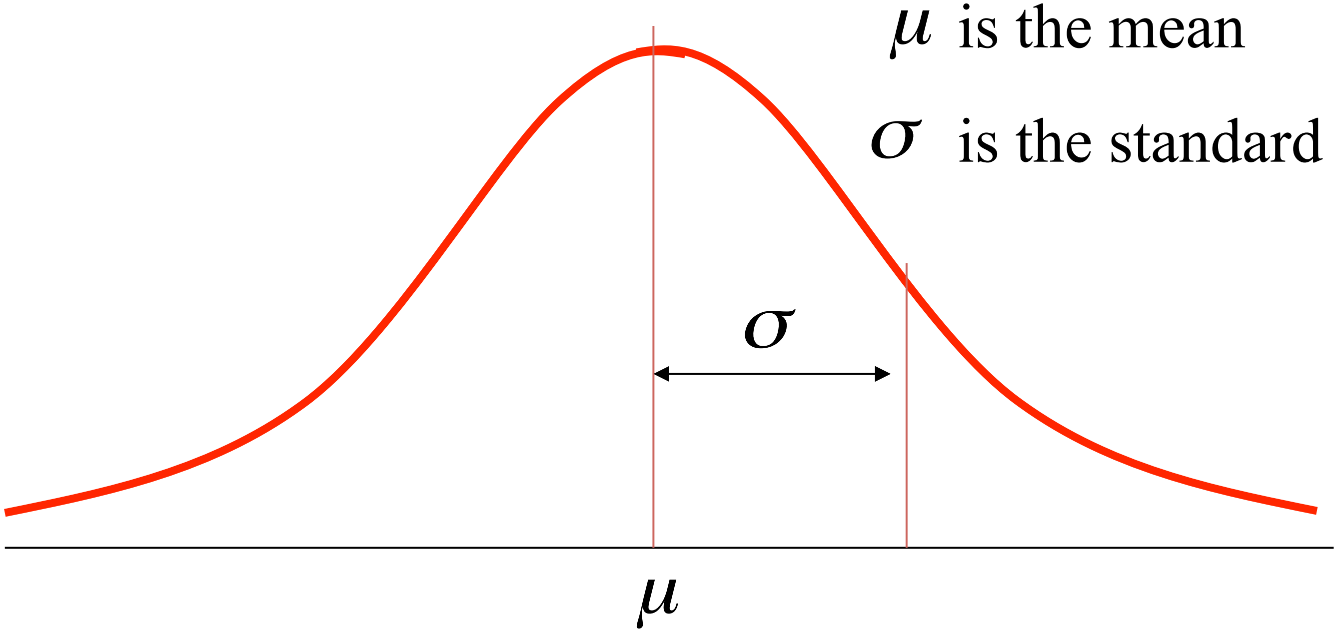
- The mean, median, and mode are the same.

The height of a normal density curve at any point x is given by

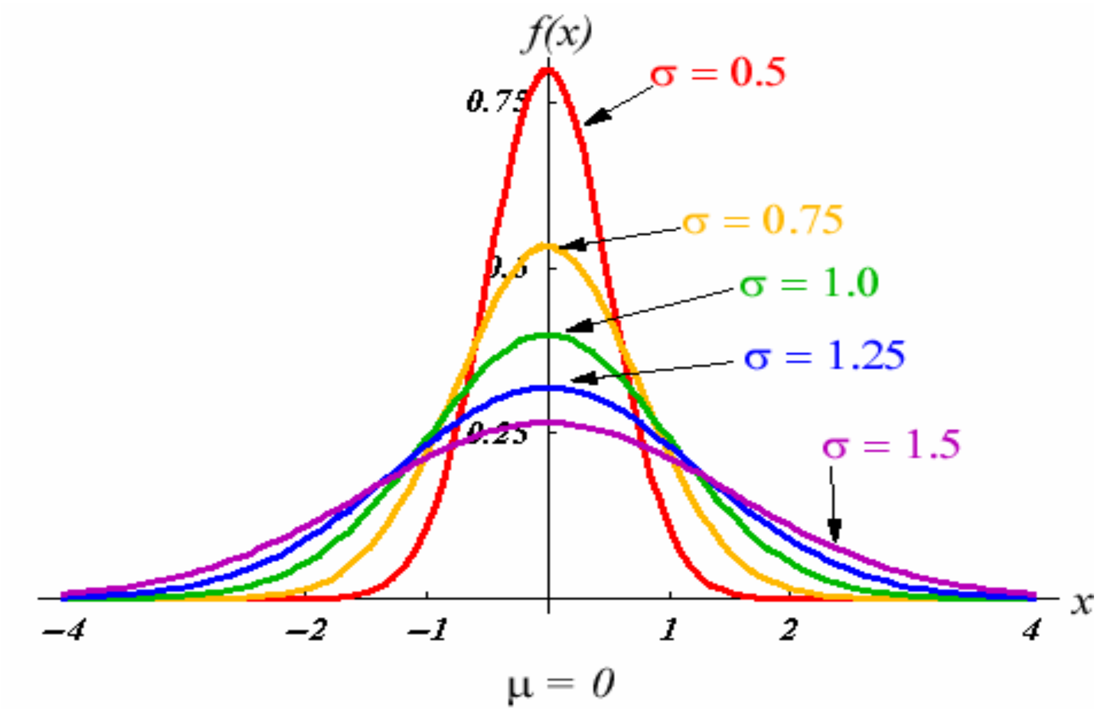
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ is the mean

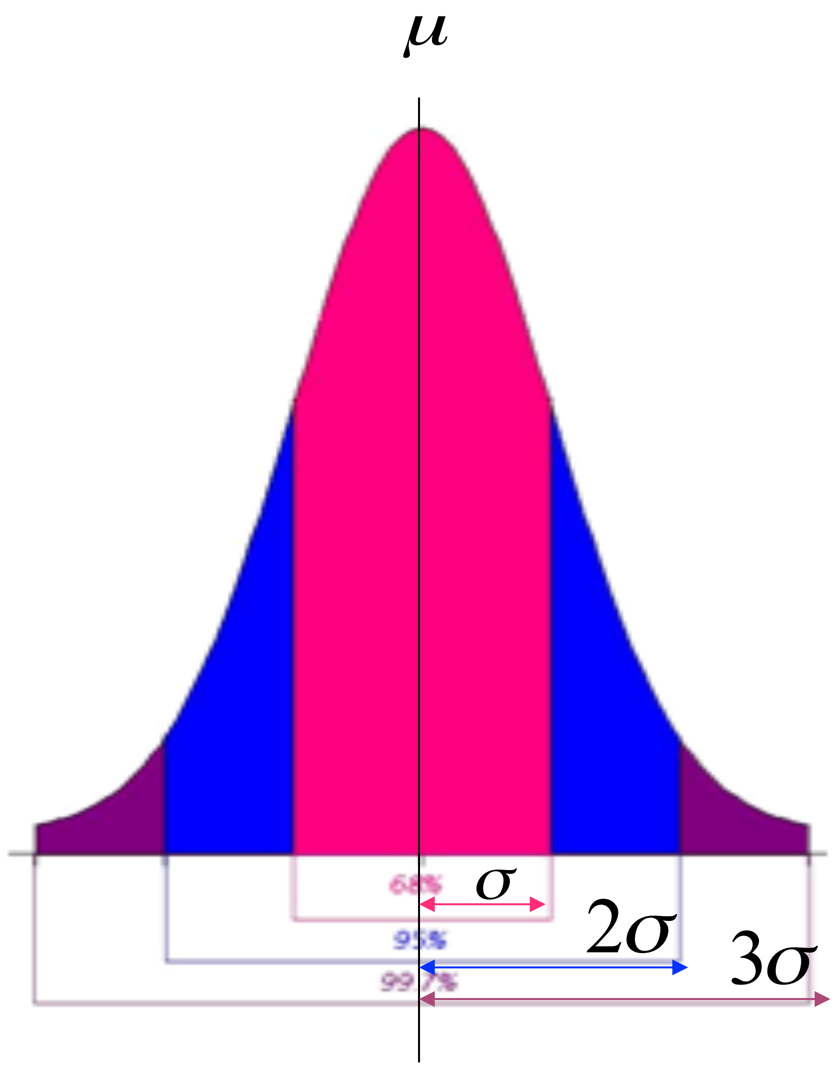
σ is the standard deviation



Example: The normal distribution is the most important distribution in Statistics. Typical normal curves with different sigma (standard deviation) values are shown below.



The 68-95-99.7 Rule for General Normal Dist.



Examples with approximate Normal distributions

- Heights
- Weights
- IQ scores
- Standardized test scores
- Body temperatures
- Repeated measurements of the same quantity
- …

Standardizing and *z*-Scores

- An observation x comes from a distribution with mean μ and standard deviation σ
- The standardized value of x is defined as

$$z = \frac{x - \mu}{\sigma},$$

which is also called a **z-score.**

- A z-score indicates how many standard deviations the original observation is away from the mean, and in which direction.

- Mean and S.D. of the distribution of z?

Effects of Standardizing

- Standardizing is a linear transformation. What are *a* and *b*?
- Effects on shape, center and spread.
- The standardized values for any distribution always have mean 0 and standard deviation 1.
- Linear transformation: normal into normal.

The standard normal distribution

- The **standard normal distribution** is the normal dist. with mean 0 and standard deviation 1, denoted as $N(0,1)$.

- $N(0,1)$ can be treated as a baseline.

- Any normal distribution can be related to $N(0,1)$ by a linear transformation.

- Z: $N(0,1)$, what is the distribution for $X=a+bZ$?

Table: The Standard Normal Table

- **Table A** is a table of areas under the standard normal density curve. The table entry for each value *z* is the area under the curve to the left of *z*.

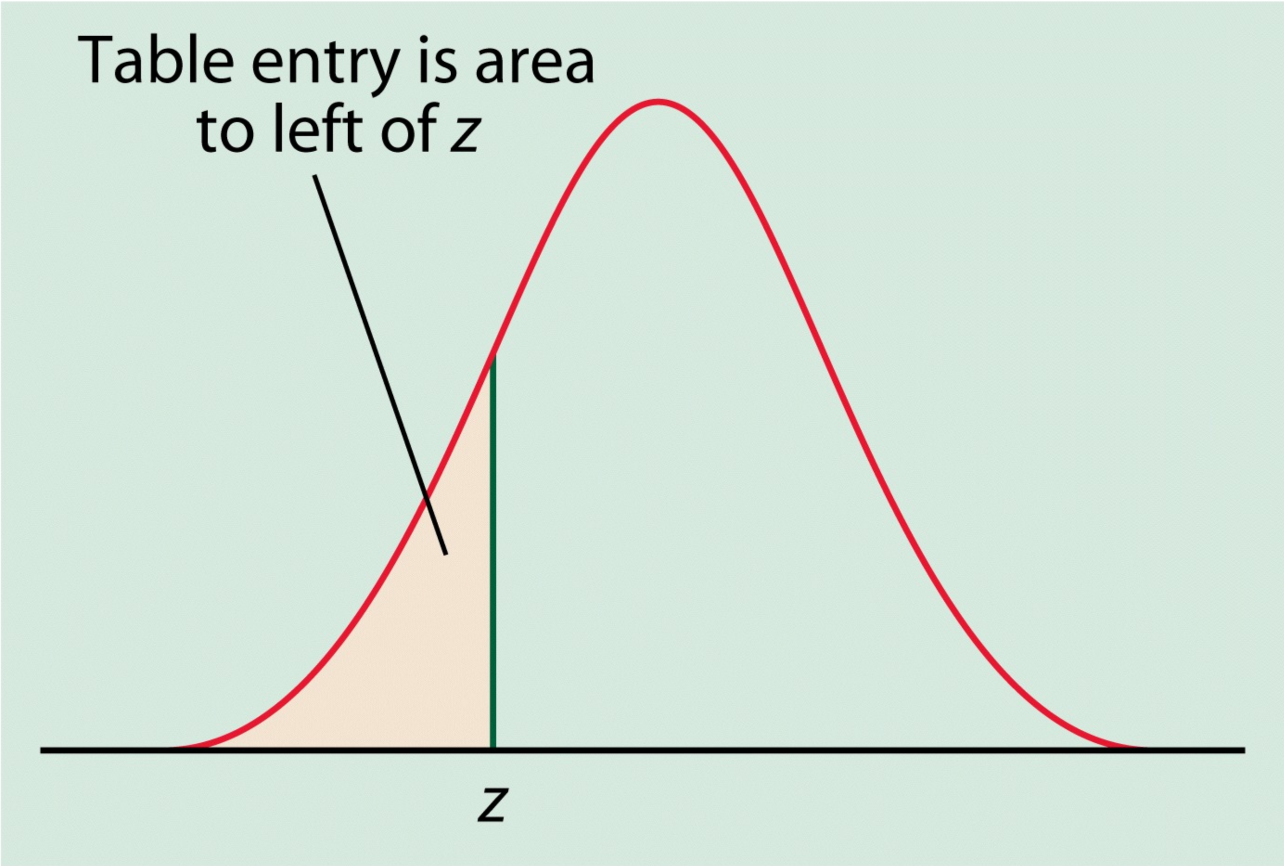
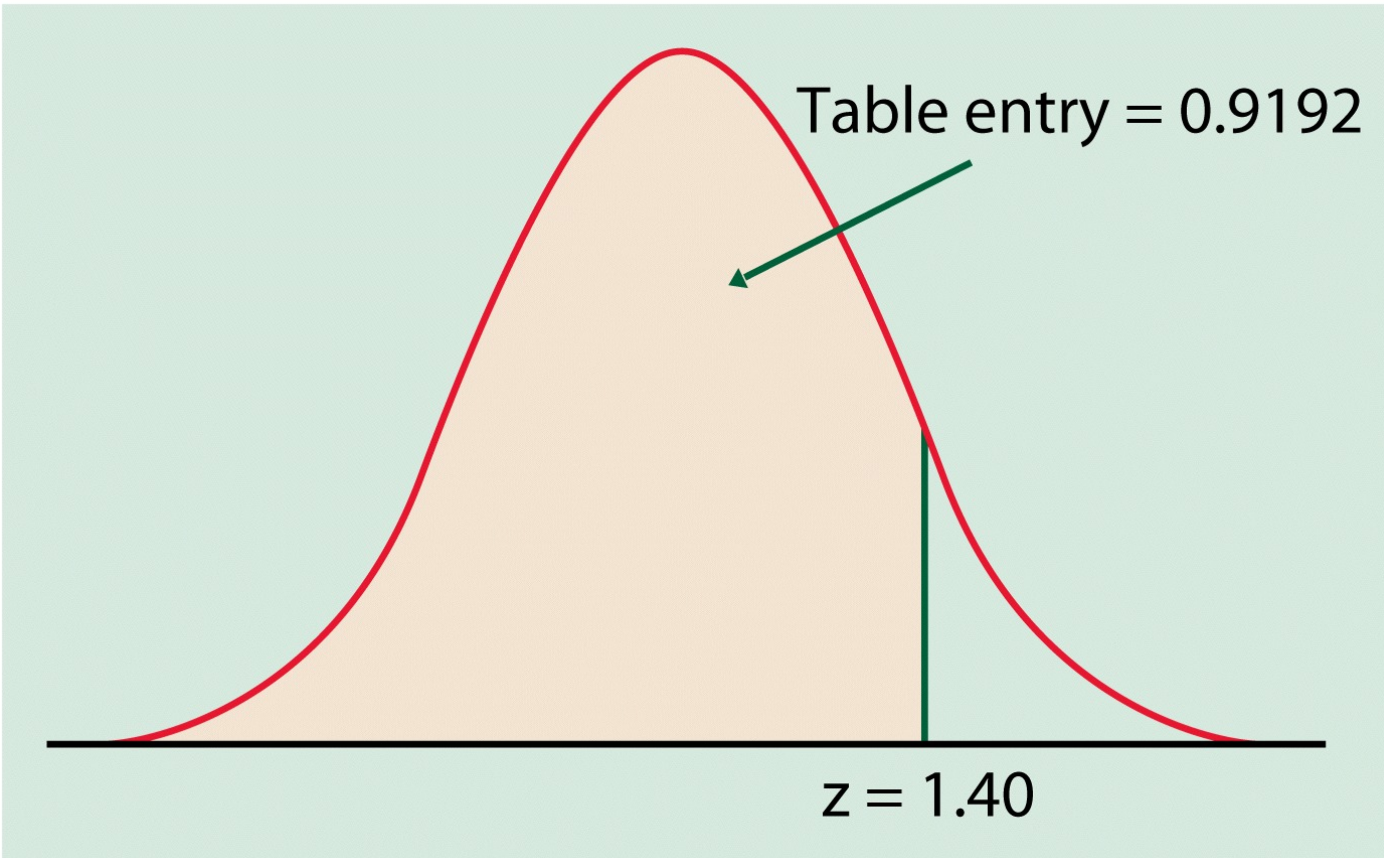


Table: The Standard Normal Table

- **Table A** can be used to find the proportion of observations of a variable which fall to the left of a specific value *z* if the variable follows a normal distribution.



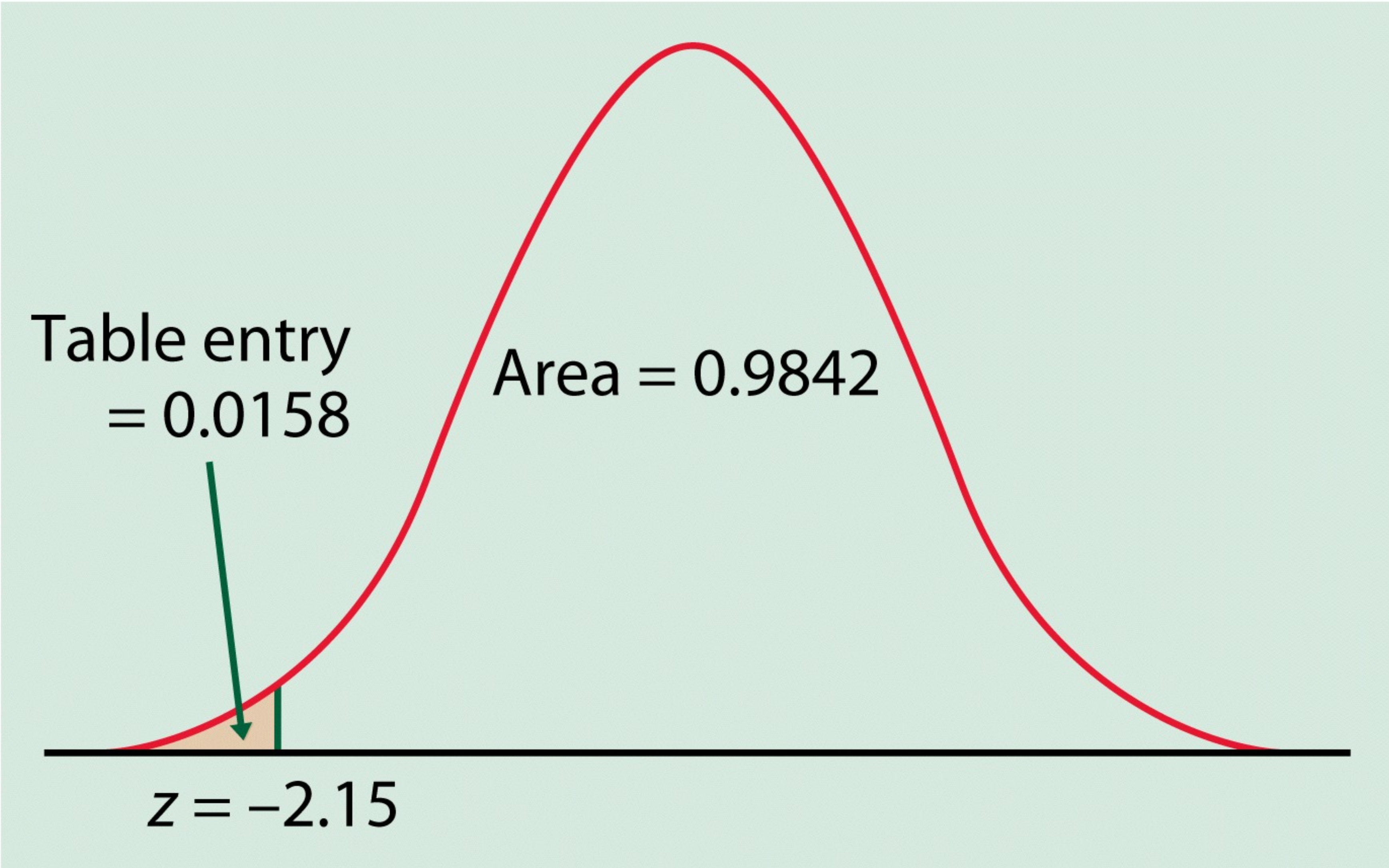


Table entry
= 0.0158
Area = 0.9842
z = −2.15

Example

- According to well-documented norms, the distribution of gestation time is approximately normal with mean 266 days and SD 16 days.
- What percent of babies have a gestation time greater or equal to 310 days (10 months and 5 days)?
- What about within a week of the due date?

Inverse problem

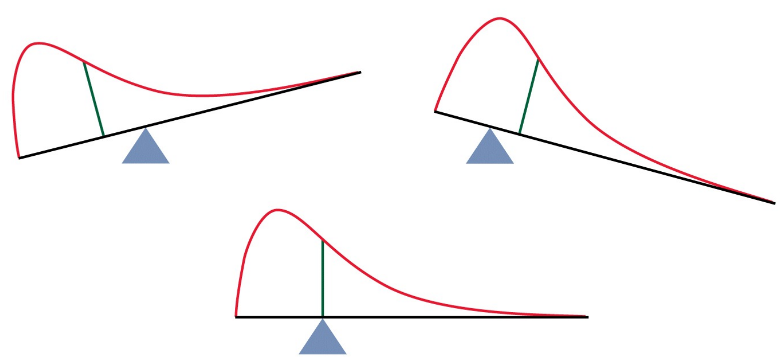
- Scores on the SAT verbal test in recent years follow approximately the $N(505, 110)$ distribution.
- How high must a student score in order to be placed in the top 10% of all students taking the SAT?

Parameters (Section 1.5)

- In general parameters are numerical summaries of the population.
- They are usually denoted by Greek letters, e.g., μ, σ, ρ, ξ, θ, η.
- Describe both univariate and multivariate populations

Parameters

- Mean
  - If the population consists of equally likely numbers ($y_1,\ldots,y_N$) then $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$ [Notice the upper case]
  - If not equally likely $\mu = \sum_{i=1}^{N} Y_i p_i$ where p_i is the probability. (Explain on a picture)



Parameters

- Standard Deviation
  - If the population consists of equally likely numbers ($Y_1,\ldots,Y_N$) then

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2}$$

  - If not equally likely

$$\sigma = \sqrt{\sum_{i=1}^{N}(Y_i - \mu)^2 p_i}$$

- Variance = (standard deviation)^2

Multivariate Distributions

- Schematic representation

	k Measurements on Each Item			
Items	1	2	$\cdots$	k
1	X_{11}	X_{12}	$\cdots$	X_{1k}
2	X_{21}	X_{22}	$\cdots$	X_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	X_{I1}	X_{I2}	$\cdots$	X_{Ik}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_{N1}	X_{N2}	$\cdots$	X_{Nk}
Mean	μ_1	μ_2	$\cdots$	μ_k
Standard deviation	σ_1	σ_2	$\cdots$	σ_k

- Each measurement has its own mean and s.d.
- Each pair of measurements has a correlation

Multivariate Distributions

- Linear relationship between measurements are described by correlations.
  - If items are equally likely

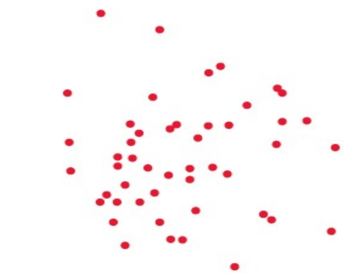
$$\rho_{Y,X} = \frac{\sum_{I=1}^{N}(Y_I - \mu_Y)(X_I - \mu_X)}{\sqrt{\left[\sum_{I=1}^{N}(Y_I - \mu_Y)^2\right]\left[\sum_{I=1}^{N}(X_I - \mu_X)^2\right]}}$$

  - Items are not equally likely

$$\rho_{Y,X} = \frac{\sum_{i=1}^{N}(Y_i - \mu_Y)(X_i - \mu_X)p_i}{\sigma_X \sigma_Y}$$

- Meaning
  - Always -1≤ρ≤1, if independent ρ=0

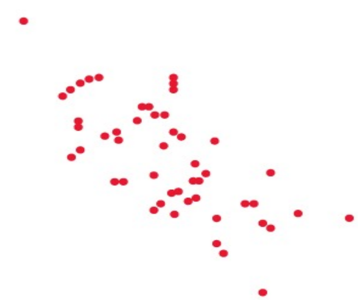
Meaning of correlation



Correlation $r = 0$

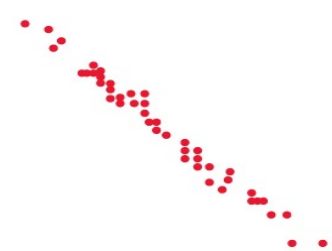
Correlation $r = -0.3$

Correlation $r = 0.5$



Correlation $r = -0.7$

Correlation $r = 0.9$



Correlation $r = -0.99$

Samples (Section 1.6)

- Populations are often unavailable to study directly.
- Solution:
  - Select a subsets of items (y_1,…,y_n) [Notice the lower case]
  - If the selection is done through a well defined random procedure, a *measure of uncertainty* can be calculated.
- Possible selection procedures
  - *Simple random sampling*
  - Other methods: stratified sampling, probability proportional to size sampling, …

Simple random sampling (SRS)

- In SRS, all the samples with the same size are equally likely to be chosen.
  - Eliminate bias in a sample.
- To conduct a SRS
  - give each unit a number
  - randomly select the sample numbers. Use a random digits table, or a software package.
- Suppose we have 100 students, need to choose 4 to answer a set of questions.
  - randomly select four number between 00 and 99
- If there are N units in the population and we want to select n we have possible samples

$$H = \binom{N}{n} = \frac{N \times (N-1) \times \cdots \times (N-n+1)}{1 \times 2 \times \cdots \times n}$$

Inferential Procedures

- Once the sample is selected, inference about the sample is performed
- Types of inference
  - Point Estimation
  - Confidence Intervals
  - Hypothesis Testing