Probability and Distributions

Part 2

Outline

- Populations and samples
- Statistical inference
- Coin-tossing example
- Probability
- Probability distributions
- Calculating probabilities for a normal distribution

Populations and Samples

- A population is a collection of observations
  - A **parameter** is a descriptive measure computed from the data of an entire population
- A sample is a part or subset of a population
  - A **statistic** is a descriptive measure computed from the data of a sample

Statistical Inference

- Statistical inference is the attempt to make a conclusion about a **population** based on a **sample**
- Statistical inference uses **data** to surmise what is true or likely to be true
- Statistical inferences requires the tools of **probability** to make an inference about the truth (unknown) based on the data

Uses of Statistical Inference

- Does tobacco cause lung cancer?
- Is there a greater response rate associated with Drug A than with Drugs B or C?
- Does Vitamin A reduce childhood mortality?
- What proportion of women develop lung cancer?

Coin-Tossing Example

- Truth
  - H_0 : coin has 2 tails
  - H_1 : coin has 1 head, 1 tail
  - H_2 : coin has 2 heads
- Probability model
  - Pr(heads given H_0 : 2 tails) =0
  - Pr(heads given H_1 : 1 head, 1 tail) = 1/2
  - Pr(heads given H_2 : 2 heads) = 1

Coin-Tossing Example

- Data: Flip a coin and observe a head
- Truth and Inference
  - Truth: H_0 Inference: Can't be
  - Truth: H_1 Inference: Possible
  - Truth: H_2 Inference: More likely
- Evidence is relative; data favor one hypothesis relative to another

Probability

- Probability provides a measure of the uncertainty associated with the occurrence of events or outcomes.
- A probability is always non-negative.
- Important criteria for outcomes:
  - Mutually exclusive events
  - Statistically independent events

Example: 2x2 Table

- Mutually exclusive events?

Gender	**Diseased**	**Not Diseased**	**Total**
Male	30	10	40
Female	50	10	60
Total	80	20	100

- P(Disease in Males) = 30/40
- P(Disease in Females) = 50/60

Another example: 2x2 Table

- Screening test for disease

Test	Diseased	Not Diseased	Total
Positive	30	10	40
Negative	50	10	60
Total	80	20	100

Another example: 2x2 Table

-
- Sensitivity = P(+ in Disease) = 85/100=0.85
- Specificity = P(- in No disease) = 180/200=0.90

Test Result	Disease	No Disease	Total
+	85	20	105
_	15	180	195
Total	100	200	300

Coin-Tossing Example

- 4 outcomes for flipping 2 coins
- For a fair coin, P(H) = P(T) = 1/2
- Two independent flips of the coin
- The probability of each outcome = 1/4
- Outcomes are mutually exclusive
  - 2 tails (TT)
  - 1 head, 1 tail (HT, TH)
  - 2 heads (HH)

Coin-Tossing Example

- Relative Frequency Table

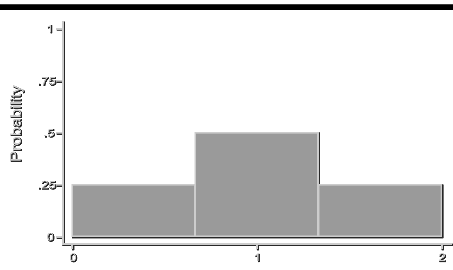
Outcome	Frequency	Probability
TT	1	1/4
HT	1	1/4
TH	1	1/4
HH	1	1/4
Total	**4**	**1.0**

Coin-Tossing Example

- Probability Distribution of the Number of Heads (X)

No. of Heads (x)	Frequency	P(X=x)
0 (TT)	1	1/4
1 (TH or HT)	2	2/4
2 (HH)	1	1/4
Total	**4**	**1.0**

Coin-Tossing Example



Coin-Tossing Example

- The probability of obtaining at least one head when flipping 2 coins =
  P(TH) + P(HT) + P(HH) = 3/4
- The probability of no heads when flipping 2 coins =
  P(TT) = 1/4 = 1 - P(at least one head)

Probability Distributions

- Theoretical probability distributions are used to describe variables of interest
  - Discrete distributions
    - Binomial
    - Poisson
  - Continuous distributions
    - Normal (Gaussian)
    - Exponential

Discrete Distributions

- Binomial distribution
  - Two possible outcomes for each of n possible "trials" (e.g., heads/tails, dead/alive, diseased/not diseased)
  - Underlies much of statistical applications to epidemiology; basic model for logistic regression methods
- Poisson distribution
  - Uses counts of events or rates

Discrete Distributions - Applications

- Binomial distribution
  - What is the probability of 2 negative mammograms in 3 tests for a woman who has breast cancer?
- Poisson distribution
  - Is the rate of memory loss higher among Gulf War veterans than the general US public?

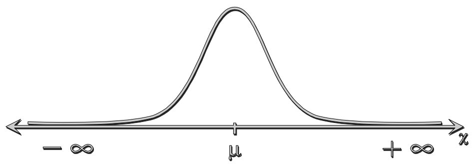
Continuous Distributions

- Normal distribution
  - Bell-shaped curve
  - Takes on values between $-\infty$ and $+\infty$
  - Symmetrical about its mean μ
  - The standard deviation of the distribution is σ
  - Mean = median = mode
  - Area under the curve = 1

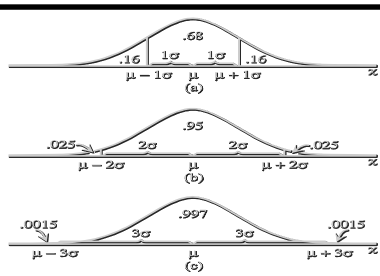
Continuous Distributions - Applications

- What is the probability of very low birth weight (< 1750 gm) infants in a certain population?
- Do mean serum cholesterol levels differ between men in a diet program and men in a diet and exercise program?

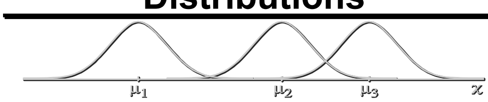
Normal Distribution



Area under a Normal Curve

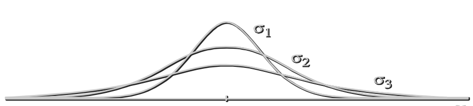


Different Normal Distributions



$\mu_1 < \mu_2 < \mu_3$

Three Normal Distributions with Different Means



$\sigma_1 < \sigma_2 < \sigma_3$

Three Normal Distributions with Different Standard Deviations

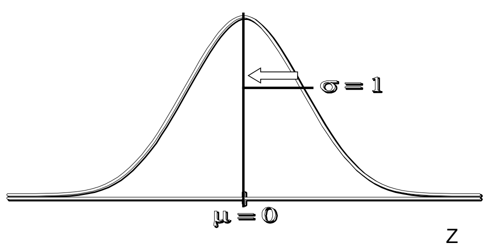
Standard Normal Distribution

- Every normal probability distribution can be transformed to a standard normal distribution using

$$Z = \frac{X - \mu}{\sigma}$$

- The standard normal distribution has mean μ = 0 and standard deviation σ = 1.
- Tables and computer programs available

Standard Normal Distribution



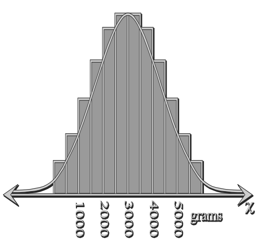
Some Common Z Values

- Z=1.96 $\Rightarrow$ P(Z > 1.96) =0.025
  - one-tailed probability = 0.025
  - two-tailed probability = 0.05
- Z=1.645 $\Rightarrow$ P(Z > 1.645) =0.05
  - one-tailed probability = 0.05
  - two-tailed probability = 0.10

Interpretation

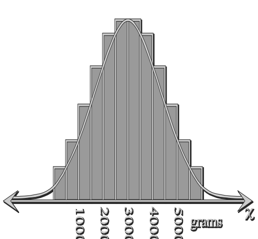
- Z > 1.96 or Z < -1.96 (values falling 2 or more standard deviations from the mean) are "unlikely" to occur just by chance alone (occur < 5% of the time)
- Z values falling within ± 1.96 standard deviations from the mean are "likely" to occur just by chance alone (occur > 5% of the time)

Example: Infant Birth Weights



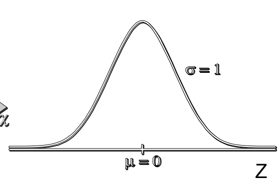
- Continuous data
- Mean = median = mode = 3000g = μ
- Standard deviation = 1000 g = σ
- Area under the curve represents the probability of infants with birth weights between certain values

Example: Infant Birth Weights

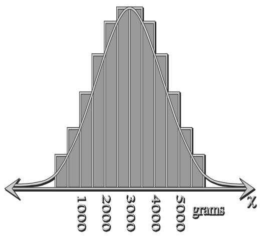


- Standard normal distribution

$$Z = \frac{X - \mu}{\sigma}$$



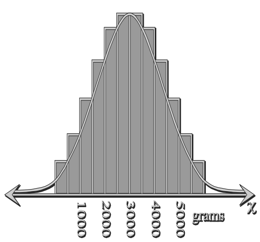
Calculating Probabilities



- What is the probability of an infant weighing more than 5000 g?

31

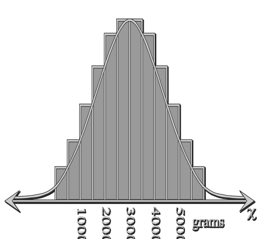
Calculating Probabilities



- What is the probability of an infant weighing less than 1000 g?

32

Calculating Probabilities



- What is the probability of an infant weighing less than 1000 g or more than 5000 g?

33

Summary

- Statistical inference about a population is based on a sample
- Statistical inference requires the tools of probability
- Probability provides a measure of certainty (or uncertainty)
- Probability distributions are useful to describe phenomena of interest