

UCLA STAT 13
**Introduction to Statistical Methods for the
 Life and Health Sciences**

Instructor: Ivo Dinov,
 Asst. Prof. of Statistics and Neurology

Teaching Assistants:
 Fred Phoa, Kirsten Johnson, Ming Zheng & Matilda Hsieh

University of California, Los Angeles, Fall 2005
http://www.stat.ucla.edu/~dinov/courses_students.html

Slide 1 Stat 13, UCLA, Ivo Dinov

Probability

- Probability is important to statistics because:
 - study results can be influenced by variation
 - it provides theoretical groundwork for statistical inference
- $0 \leq P(A) \leq 1$
 - In English please: the probability of event A **must** be between zero and one.
 - Note: $P(A) = \Pr(A)$

Slide 2 Stat 13, UCLA, Ivo Dinov

Random Sampling

- A simple random sample of n items is a sample in which:
 - every member of the population has an equal chance of being selected.
 - the members of the sample are chosen independently.

Slide 3 Stat 13, UCLA, Ivo Dinov

Random Sampling

Example: Consider our class as the population under study. If we select a sample of size 5, each possible sample of size 5 must have the same chance of being selected.

- When a sample is chosen randomly it is the process of selection that is random.
- How could we randomly select five members from this class randomly?

Slide 4 Stat 13, UCLA, Ivo Dinov

Random Sampling

- Random Number Table (e.g., Table 1 in text)
- Random Number generator on a computer (e.g., www.socr.ucla.edu SOCR Modeler → Random Number Generation)
- Which one is the best?
- Example (cont'): Let's randomly select five students from this class using the table and the computer.

Slide 5 Stat 13, UCLA, Ivo Dinov

Random Sampling

Table Method (p. 670 in book):

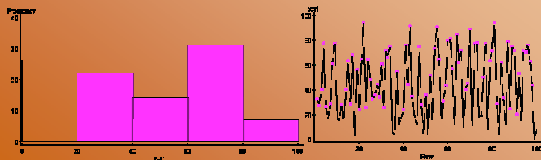
1. Randomly assign id's to each member in the population (1 - n)
2. Choose a place to start in table (close eyes)
3. Start with the first number (must have the same number of digits as n), this is the first member of the sample.
4. Work left, right, up or down, just stay consistent.
5. Choose the next number (must have the same number of digits as n), this is the second member of the sample.
6. Repeat step 5 until all members are selected. If a number is repeated or not possible move to the next following your algorithm.

Slide 6 Stat 13, UCLA, Ivo Dinov

Random Sampling

Computer Method:

1. http://socr.stat.ucla.edu/htmls/SOCR_Modeler.html
2. Data Generation → Discrete Uniform Distribution.
3. Histogram plot (left) and Raw Data index Plot (Right)



Slide 7 Stat 13, UCLA, Jon Dineen

Key Issue

- How representative of the population is the sample likely to be?
 - The sample wont exactly resemble the population, there will be some chance variation. This discrepancy is called "chance error due to sampling".
- **Definition:** *Sampling bias* is non-randomness that refers to some members having a tendency to be selected more readily than others.
 - When the sample is biased the statistics turn out to be poor estimates.

Slide 8 Stat 13, UCLA, Jon Dineen

Key Issue

Example: Suppose a weight loss clinic is interested in studying the effects of a new diet proposed by one of it researchers. It decides to advertise in the LA Times for participants to come be part of the study.

Example: Suppose a lake is to be studied for toxic emissions from a nearby power plant. The samples that were obtained came from the portion of the lake that was the closest possible location to the plant.

Slide 9 Stat 13, UCLA, Jon Dineen

Let's Make a Deal Paradox – aka, Monty Hall 3-door problem



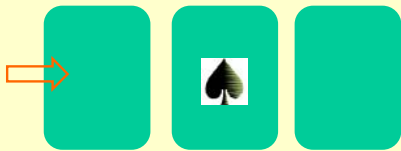
- This paradox is related to a popular television show in the 1970's. In the show, a contestant was given a choice of **three doors/cards** of which one contained a prize (**diamond**). The other two doors contained gag gifts like a chicken or a donkey (clubs).



Slide 10 Stat 13, UCLA, Jon Dineen

Let's Make a Deal Paradox.

- After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?



1. Pick One card
2. Show one Club Card
3. Change 1st pick?

Slide 11 Stat 13, UCLA, Jon Dineen

Let's Make a Deal Paradox.

- The *intuition* of most people tells them that each of the doors, the chosen door and the unchosen door, are equally likely to contain the prize so that there is a **50-50 chance** of winning with either selection? This, however, is **not the case**.
- The **probability of winning by using the switching technique is 2/3**, while the odds of winning by not switching is 1/3. The easiest way to explain this is as follows:

Slide 12 Stat 13, UCLA, Jon Dineen

Let's Make a Deal Paradox.

- The probability of picking the wrong door in the initial stage of the game is $2/3$.
- If the contestant picks the wrong door initially, the host must reveal the remaining empty door in the second stage of the game. Thus, if the contestant switches after picking the wrong door initially, the contestant will win the prize.
- The probability of winning by switching then reduces to the probability of picking the wrong door in the initial stage which is clearly $2/3$.

● **Demos:**

- file:///C:/Ivo.dir/UCLA_Classes/Applets.dir/SOCR/Prototype1.1/classes/TestExperiment.html
- C:/Ivo.dir/UCLA_Classes/Applets.dir/StatGames.exe

Slide 13 Stat 13, UCLA, Ivo Dinner

Long run behavior of coin tossing

Number of tosses

Figure 4.1.1 Proportion of heads versus number of tosses for John Kerrich's coin tossing experiment.

Slide 14 Stat 13, UCLA, Ivo Dinner

Definitions ...

- The **law of averages** about the behavior of coin tosses – the **relative proportion** (relative frequency) of heads-to-tails in a coin toss experiment becomes more and **more stable** as the **number of tosses increases**. The **law of averages** applies to **relative frequencies not absolute counts** of #H and #T.
- Two widely held **misconceptions** about what the **law of averages** about coin tosses:
 - Differences between the actual numbers of heads & tails becomes more and more variable with increase of the number of tosses – a seq. of 10 heads doesn't increase the chance of a tail on the next trial.
 - Coin toss results are **fair**, but behavior is still **unpredictable**.

Slide 15 Stat 13, UCLA, Ivo Dinner

Coin Toss Models

- Is the **coin tossing model** adequate for describing the **sex order** of children in families?
 - This is a rough model which is not exact. In most countries **rates of B/G is different**; form 48% ... 52%, usually. Birth rates of boys in some places are higher than girls, however, female population seems to be about 51%.
 - **Independence**, if a second child is born the chance it has the same gender (as the first child) is slightly bigger.

Slide 16 Stat 13, UCLA, Ivo Dinner

Data from a "random" draw

Month of the year

Figure 4.3.1 Average lottery numbers by month. Replotted from data in Fienberg [1971].

Slide 17 Stat 13, UCLA, Ivo Dinner

Types of Probability

- Probability models have two essential components (**sample space**, the space of all possible outcomes from an experiment; and a list of **probabilities** for each event in the sample space). Where do the **outcomes** and the **probabilities** come from?
- **Probabilities from models** – say mathematical/physical description of the sample space and the chance of each event. Construct a fair die tossing game.
- **Probabilities from data** – data observations determine our probability distribution. Say we toss a coin 100 times and the observed Head/Tail counts are used as probabilities.
- **Subjective Probabilities** – combining data and psychological factors to design a reasonable probability table (e.g., gambling, stock market).

Slide 18 Stat 13, UCLA, Ivo Dinner

Sample Spaces and Probabilities

- When the relative frequency of an event in the past is used to estimate the probability that it will occur in the future, what assumption is being made?
 - The underlying process is stable over time;
 - Our relative frequencies must be taken from large numbers for us to have confidence in them as probabilities.
- All statisticians agree about how probabilities are to be combined and manipulated (in math terms), however, not all agree what probabilities should be associated with for a particular real-world event.
- When a weather forecaster says that there is a 70% chance of rain tomorrow, what do you think this statement means? (Based on our past knowledge, according to the barometric pressure, temperature, etc. of the conditions we expect tomorrow, 70% of the time it did rain under such conditions.)

Slide 19 Stat 13, UCLA, Jon Dinger

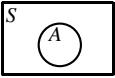
Sample spaces and events

- A **sample space**, S , for a random experiment is the set of all possible outcomes of the experiment.
- An **event** is a collection of outcomes.
- An event **occurs** if any outcome making up that event occurs.

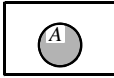
Slide 20 Stat 13, UCLA, Jon Dinger

The complement of an event

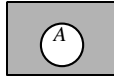
- The **complement** of an event A , denoted \bar{A} , occurs if and only if A does not occur.



(a) Sample space containing event A



(b) Event A shaded




(c) \bar{A} shaded

Figure 4.4.1 An event A in the sample space S .


Slide 21 Stat 13, UCLA, Jon Dinger

Combining events – all statisticians agree on


- “**A or B**” contains all outcomes in A or B (or both).
- “**A and B**” contains all outcomes which are in both A and B .




(a) Events A and B



(b) “ A or B ” shaded



(c) “ A and B ” shaded



(d) Mutually exclusive events

Figure 4.4.2 Two events.

From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Mutually exclusive events cannot occur at the same time.

Slide 22 Stat 13, UCLA, Jon Dinger

Probability distributions

- Probabilities always lie between 0 and 1 and they sum up to 1 (across all simple events).
- $pr(A)$ can be obtained by adding up the probabilities of all the outcomes in A .

$$pr(A) = \sum_{\substack{E \text{ outcome} \\ \text{in event } A}} pr(E)$$

Slide 23 Stat 13, UCLA, Jon Dinger

Job losses in the US

TABLE 4.4.1 Job Losses in the US (in thousands) for 1987 to 1991

| | Reason for Job Loss | | | Total |
|--------|------------------------|------------|--------------------|-------|
| | Workplace moved/closed | Slack work | Position abolished | |
| Male | 1,703 | 1,196 | 548 | 3,447 |
| Female | 1,210 | 564 | 363 | 2,137 |
| Total | 2,913 | 1,760 | 911 | 5,584 |

Slide 24 Stat 13, UCLA, Jon Dinger

Job losses cont.

| | Workplace moved/closed | Slack work | Position abolished | Total |
|--------|---------------------------|------------|-----------------------|-------|
| Male | 1,703 | 1,196 | 548 | 3,447 |
| Female | 1,210 | 564 | 363 | 2,137 |
| Total | 2,913 | 1,760 | 911 | 5,584 |

TABLE 4.4.2 Proportions of Job Losses from Table 4.4.1

| | Reason for Job Loss | | | Row totals |
|---------------|---------------------------|------------|-----------------------|---------------|
| | Workplace moved/closed | Slack work | Position abolished | |
| Male | .305 | .214 | .098 | .617 |
| Female | .217 | .101 | .065 | .383 |
| Column totals | .552 | .315 | .163 | 1.000 |

Slide 25 Stat 13, UCLA, Jon Dinger

- Review**
- What is a **sample space**? What are the **two essential criteria** that must be satisfied by a possible sample space? (**completeness** – every outcome is represented; and **uniqueness** – no outcome is represented more than once.)
 - What is an **event**? (collection of outcomes)
 - If A is an event, what do we mean by its complement, \bar{A} ? When does \bar{A} occur?
 - If A and B are events, when does A or B occur? When does A and B occur?
- Slide 26 Stat 13, UCLA, Jon Dinger

- Properties of probability distributions**
- A sequence of number $\{p_1, p_2, p_3, \dots, p_n\}$ is a **probability distribution** for a sample space $S = \{s_1, s_2, s_3, \dots, s_n\}$, if $\text{pr}(s_k) = p_k$, for each $1 \leq k \leq n$. The two essential **properties of a probability distribution** p_1, p_2, \dots, p_n ?

$$p_i \geq 0; \sum p_i = 1$$
 - How do we get the probability of an event from the probabilities of outcomes that make up that event?
 - If all outcomes are **distinct & equally likely**, how do we calculate $\text{pr}(A)$? If $A = \{a_1, a_2, a_3, \dots, a_9\}$ and $\text{pr}(a_1) = \text{pr}(a_2) = \dots = \text{pr}(a_9) = p$; then

$$\text{pr}(A) = 9 \times \text{pr}(a_i) = 9p.$$
- Slide 27 Stat 13, UCLA, Jon Dinger

- Example of probability distributions**
- Tossing a coin twice. **Sample space** $S = \{HH, HT, TH, TT\}$, for a fair coin each outcome is equally likely, so the probabilities of the 4 possible outcomes should be identical, p . Since, $p(HH) = p(HT) = p(TH) = p(TT) = p$ and

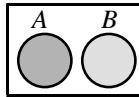
$$p_i \geq 0; \sum p_i = 1$$
 - $p = 1/4 = 0.25$.
- Slide 28 Stat 13, UCLA, Jon Dinger

- Proportion vs. Probability**
- How do the concepts of a **proportion** and a **probability differ**? A **proportion** is a **partial description** of a real population. The **probabilities** give us the **chance** of something happening in a random experiment. Sometimes, **proportions** are **identical** to **probabilities** (e.g., in a real population under the experiment **choose-a-unit-at-random**).
 - See the **two-way table of counts (contingency table)** on Table 4.4.1, slide 19. E.g., **choose-a-person-at-random** from the ones laid off, and compute the chance that the person would be a **male**, laid off due to **position-closing**. We can apply the same rules for manipulating probabilities to proportions, in the case where these two are identical.
- Slide 29 Stat 13, UCLA, Jon Dinger

Rules for manipulating Probability Distributions

For mutually exclusive events,

$$\text{pr}(A \text{ or } B) = \text{pr}(A) + \text{pr}(B)$$



From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Slide 30 Stat 13, UCLA, Jon Dinger

Descriptive Table

| | | | | | | | | |
|--------------------------|-------|-----|-------|-----------------|-----------|-------------------------------------|---|----------------------|
| pr(Wild in and Seber in) | Seber | | Total | Algebraic Table | | | | |
| | In | Out | | | B | \bar{B} | Total | |
| Wild | In | 0.5 | ? | 0.7 | A | $\text{pr}(A \text{ and } B)$ | $\text{pr}(A \text{ and } \bar{B})$ | $\text{pr}(A)$ |
| | Out | ? | ? | ? | \bar{A} | $\text{pr}(\bar{A} \text{ and } B)$ | $\text{pr}(\bar{A} \text{ and } \bar{B})$ | $\text{pr}(\bar{A})$ |
| | Total | 0.6 | ? | 1.00 | Total | $\text{pr}(B)$ | $\text{pr}(\bar{B})$ | 1.00 |

Availability of the Textbook authors to students

| | | | | | | | | | | | |
|----|---|------|----|----|------|----|----|------|----|----|------|
| .5 | ? | .7 | .5 | ? | .7 | .5 | .2 | .7 | .5 | .2 | .7 |
| ? | ? | ? | ? | ? | ? | ? | ? | .3 | ? | ? | .3 |
| .6 | ? | 1.00 | .6 | .4 | 1.00 | .6 | .4 | 1.00 | .6 | .4 | 1.00 |

TABLE 4.5.1
Completed Probability Table

| | | | |
|-------|-------|-----|-------|
| | Seber | | Total |
| Wild | In | Out | |
| In | .5 | .2 | .7 |
| Out | .1 | .2 | .3 |
| Total | .6 | .4 | 1.0 |

Slide 31 Stat 13, UCLA, Jon Dinger

Unmarried couples

Select an unmarried couple *at random* – the table proportions give us the probabilities of the events defined in the row/column titles.

TABLE 4.5.2 Proportions of Unmarried Male-Female Couples Sharing Household in the US, 1991

| | | | | | | |
|-------|------------------|----------|---------|------------------|-------|------|
| | Female | | | | Total | |
| | Never Married | Divorced | Widowed | Married to other | | |
| Male | Never Married | .401 | .111 | .017 | .025 | .554 |
| | Divorced | .117 | .195 | .024 | .017 | .353 |
| | Widowed | .006 | .008 | .016 | .001 | .031 |
| | Married to other | .021 | .022 | .003 | .016 | .062 |
| Total | .545 | .336 | .060 | .059 | 1.000 | |

Slide 32 Stat 13, UCLA, Jon Dinger

Review

- If A and B are **mutually exclusive**, what is the probability that **both occur**? (\cap) What is the probability that **at least one occurs**? (sum of probabilities)
- If we have two or more mutually exclusive events, how do we find the probability that **at least one of them occurs**? (sum of probabilities)
- Why is it sometimes easier to compute $\text{pr}(A)$ from $\text{pr}(\bar{A}) = 1 - \text{pr}(\bar{A})$? (The **complement** of the even may be easier to find or may have a known probability. E.g., a random number between 1 and 10 is drawn. Let $A = \{a \text{ number less than or equal to } 9 \text{ appears}\}$. Find $\text{pr}(A) = 1 - \text{pr}(\bar{A})$. probability of \bar{A} is $\text{pr}(\{10 \text{ appears}\}) = 1/10 = 0.1$. Also Monty Hall 3 door example!

Slide 33 Stat 13, UCLA, Jon Dinger

Melanoma – type of skin cancer – an example of laws of conditional probabilities

TABLE 4.6.1: 400 Melanoma Patients by Type and Site

| | | | | |
|--------------------------------|---------------|-------|-------------|------------|
| | Site | | | Row Totals |
| | Head and Neck | Trunk | Extremities | |
| Type | Head and Neck | Trunk | Extremities | Row Totals |
| Hutchinson's melanomic freckle | 22 | 2 | 10 | 34 |
| Superficial | 16 | 54 | 115 | 185 |
| Nodular | 19 | 33 | 73 | 125 |
| Indeterminant | 11 | 17 | 28 | 56 |
| Column Totals | 68 | 106 | 226 | 400 |

Contingency table based on Melanoma histological type and its location

Slide 34 Stat 13, UCLA, Jon Dinger

Conditional Probability

The **conditional probability** of A occurring **given** that B occurs is given by

$$\text{pr}(A | B) = \frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)}$$

Suppose we select one out of the 400 patients in the study and we want to find the probability that the cancer is on the **extremities** given that it is of type **nodular**: $P = 73/125 = P(\text{C. on Extremities} | \text{Nodular})$

#nodular patients with cancer on extremities
#nodular patients

Slide 35 Stat 13, UCLA, Jon Dinger

Multiplication rule- what's the percentage of Israelis that are poor and Arabic?

$$\text{pr}(A \text{ and } B) = \text{pr}(A | B)\text{pr}(B) = \text{pr}(B | A)\text{pr}(A)$$

Figure 4.6.1 Illustration of the multiplication rule.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Slide 36 Stat 13, UCLA, Jon Dinger

A tree diagram for computing conditional probabilities

Suppose we draw 2 balls at random one at a time *without replacement* from an urn containing **4 black** and **3 white** balls, otherwise identical. What is the probability that the *second ball is black*? Sample Spc?

Mutually exclusive

$$P(\{2\text{-nd ball is black}\}) = P(\{2\text{-nd is black} \ \& \ \{1\text{-st is black}\}) + P(\{2\text{-nd is black} \ \& \ \{1\text{-st is white}\}) = 4/7 \times 3/6 + 4/6 \times 3/7 = 4/7.$$

Slide 39 Stat 13, UCLA, Jon Dinger

A tree diagram

Figure 4.6.2 Tree diagram for a sampling problem. Slide 40 Stat 13, UCLA, Jon Dinger

Tree diagram for poverty in Israel

Slide 41 Stat 13, UCLA, Jon Dinger

2-way table for poverty in Israel

| | | | | |
|---------|----------|------------------|------------------|-------|
| | | Ethnicity | | Total |
| | | Arabic | Jewish | |
| Poverty | Poor | $.52 \times .14$ | $.11 \times .86$ | ? |
| | Not poor | ? | ? | ? |
| | Total | .14 | .86 | 1.00 |

pr(Arabic) = .14 pr(Jewish) = .86

Figure 4.6.4 Proportions by Ethnicity and Poverty. Slide 42 Stat 13, UCLA, Jon Dinger

$$P(A \ \& \ B) = P(A \ | \ B) \times P(B),$$

$$P(A \ | \ B) = P(A \ \& \ B) / P(B)$$

$$P(A \ \& \ B) = P(B \ \& \ A) = P(B \ | \ A) \times P(A).$$

$$P(A \ | \ B) = [P(B \ | \ A) \times P(A)] / P(B).$$

2-way table for poverty in Israel cont.

| | | | | |
|---------|----------|------------------|------------------|-------|
| | | Ethnicity | | Total |
| | | Arabic | Jewish | |
| Poverty | Poor | $.52 \times .14$ | $.11 \times .86$ | ? |
| | Not poor | ? | ? | ? |
| | Total | .14 | .86 | 1.00 |

pr(Arabic) = .14 pr(Jewish) = .86

TABLE 4.6.3 Proportions by Ethnicity and Poverty

| | | | | |
|---------|----------|-----------|--------|-------|
| | | Ethnicity | | Total |
| | | Arabic | Jewish | |
| Poverty | Poor | .0728 | .0946 | .1674 |
| | Not Poor | .0672 | .7654 | .8326 |
| | Total | .14 | .86 | 1.00 |

Slide 43 Stat 13, UCLA, Jon Dinger

Conditional probabilities and 2-way tables

- Many problems involving conditional probabilities can be solved by constructing two-way tables
- This includes *reversing the order of conditioning*

$$P(A \ \& \ B) = P(A \ | \ B) \times P(B) = P(B \ | \ A) \times P(A)$$

Slide 44 Stat 13, UCLA, Jon Dinger

Classes vs. Evidence Conditioning

- **Classes:** healthy(NC), cancer
- **Evidence:** positive mammogram (pos), negative mammogram (neg)
- If a woman has a positive mammogram result, what is the probability that she has breast cancer?

$$P(\text{class} | \text{evidence}) = \frac{P(\text{evidence} | \text{class}) \times P(\text{class})}{P(\text{evidence})}$$

$$P(\text{cancer}) = 0.01$$

$$P(\text{pos} | \text{cancer}) = 0.8$$

$$P(\text{positive}) = 0.107$$

$$P(\text{cancer} | \text{pos}) = ?$$

Slide 45 Stat 13, UCLA, Jon Dineen

Proportional usage of oral contraceptives and their rates of failure

We need to complete the two-way contingency table of proportions

| | | | | | | | |
|---------|--------|---------|--------------|-----------|-----------|-----------|-------|
| | | Method | | | | | Total |
| | | Steril. | Oral Barrier | IUD | Sperm. | | |
| Outcome | Failed | 0 × .38 | .05 × .32 | .14 × .24 | .06 × .03 | .26 × .03 | ? |
| | Didn't | ? | ? | ? | ? | ? | ? |
| | Total | .38 | .32 | .24 | .03 | .03 | 1.00 |

$\text{pr}(\text{Failed and Oral}) = \text{pr}(\text{Failed} | \text{Oral}) \times \text{pr}(\text{Oral})$
 [= 5% of 32%]

$\text{pr}(\text{Failed and IUD}) = \text{pr}(\text{Failed} | \text{IUD}) \times \text{pr}(\text{IUD})$
 [= 6% of 3%]

$\text{pr}(\text{Steril.}) = .38$ $\text{pr}(\text{Barrier}) = .24$ $\text{pr}(\text{IUD}) = .03$

Slide 46 Stat 13, UCLA, Jon Dineen

Oral contraceptives cont.

| | | | | | | | |
|---------|--------|---------|--------------|-----------|-----------|-----------|-------|
| | | Method | | | | | Total |
| | | Steril. | Oral Barrier | IUD | Sperm. | | |
| Outcome | Failed | 0 × .38 | .05 × .32 | .14 × .24 | .06 × .03 | .26 × .03 | ? |
| | Didn't | ? | ? | ? | ? | ? | ? |
| | Total | .38 | .32 | .24 | .03 | .03 | 1.00 |

$\text{pr}(\text{Failed and Oral}) = \text{pr}(\text{Failed} | \text{Oral}) \times \text{pr}(\text{Oral})$
 [= 5% of 32%]

$\text{pr}(\text{Failed and IUD}) = \text{pr}(\text{Failed} | \text{IUD}) \times \text{pr}(\text{IUD})$
 [= 6% of 3%]

$\text{pr}(\text{Steril.}) = .38$ $\text{pr}(\text{Barrier}) = .24$ $\text{pr}(\text{IUD}) = .03$

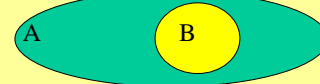
TABLE 4.6.4 Table Constructed from the Data in Example 4.6.8

| | | Method | | | | | Total |
|---------|--------|---------|--------------|-------|--------|-------|--------|
| | | Steril. | Oral Barrier | IUD | Sperm. | | |
| Outcome | Failed | 0 | .0160 | .0336 | .0018 | .0078 | .0592 |
| | Didn't | .3800 | .3040 | .2064 | .0282 | .0222 | .9408 |
| | Total | .3800 | .3200 | .2400 | .0300 | .0300 | 1.0000 |

Slide 47 Stat 13, UCLA, Jon Dineen

Remarks ...

- In $\text{pr}(A | B)$, how should the symbol “|” be read given that.
- How do we interpret the fact that: *The event A always occurs when B occurs?* What can you say about $\text{pr}(A | B)$?



- When drawing a probability tree for a particular problem, how do you know what events to use for the first fan of branches and which events to use for the subsequent branching? (at each branching stage condition on all the info available up to here. E.g., at first branching use all simple events, no prior is available. At 3-rd branching condition of the previous 2 events, etc.)

Slide 48 Stat 13, UCLA, Jon Dineen

TABLE 4.6.5 Number of Individuals Having a Given Mean Absorbance Ratio (MAR) in the ELISA for HIV Antibodies

| MAR | Healthy Donor | HIV patients |
|----------|---------------|--------------|
| <2 | 202 | 0 |
| 2 - 2.99 | 73 | 2 |
| 3 - 3.99 | 15 | 7 |
| 4 - 4.99 | 3 | 7 |
| 5 - 5.99 | 2 | 15 |
| 6 -11.99 | 2 | 36 |
| 12+ | 0 | 21 |
| Total | 297 | 88 |

Test cut-off 2 } 2 False-Negatives (FNE)
 } 2 False-positives
 Power of a test is: 1-P(FNE)= 1-P(Neg|HIV) ~ 0.976

Adapted from Weiss et al.[1985]

Slide 49 Stat 13, UCLA, Jon Dineen

HIV cont.

| | | | | |
|----------------|---------|-------------|-----------|-------|
| | | Test result | | Total |
| | | Positive | Negative | |
| Disease status | HIV | .98 × .01 | ? | .01 |
| | Not HIV | ? | .93 × .99 | .99 |
| | Total | ? | ? | 1.00 |

$\text{pr}(\text{HIV and Positive}) = \text{pr}(\text{Positive} | \text{HIV}) \times \text{pr}(\text{HIV})$
 [= 98% of 1%]

$\text{pr}(\text{Not HIV and Negative}) = \text{pr}(\text{Negative} | \text{Not HIV}) \times \text{pr}(\text{Not HIV})$
 [= 93% of 99%]

$\text{pr}(\text{HIV}) = .01$
 $\text{pr}(\text{Not HIV}) = .99$

Figure 4.6.6 Putting HIV information into the table.

From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Slide 50 Stat 13, UCLA, Jon Dineen

HIV – reconstructing the contingency table

$\text{pr}(\text{HIV and Positive}) = \text{pr}(\text{Positive}|\text{HIV}) \times \text{pr}(\text{HIV})$
 [= 98% of 1%]

$\text{pr}(\text{Not HIV and Negative}) = \text{pr}(\text{Negative}|\text{Not HIV}) \times \text{pr}(\text{Not HIV})$
 [= 93% of 99%]

| Disease status | | Test result | | Total |
|----------------|--|------------------|------------------|-------|
| | | Positive | Negative | |
| HIV | | $.98 \times .01$ | ? | .01 |
| Not HIV | | ? | $.93 \times .99$ | .99 |
| Total | | ? | ? | 1.00 |

$\text{pr}(\text{HIV}) = .01$
 $\text{pr}(\text{Not HIV}) = .99$

TABLE 4.6.6 Proportions by Disease Status and Test Result

| Disease Status | | Test Result | | Total |
|----------------|--|-------------|----------|-------|
| | | Positive | Negative | |
| HIV | | .0098 | .0002 | .01 |
| Not HIV | | .0693 | .9207 | .99 |
| Total | | .0791 | .9209 | 1.00 |

Slide 51 Stat 13, UCLA, Jon Dineen

Proportions of HIV infections by country

TABLE 4.6.7 Proportions Infected with HIV

| Country | No. AIDS Cases | Population (millions) | pr(HIV) | Having Test pr(HIV Positive) |
|----------------|----------------|-----------------------|---------|----------------------------------|
| United States | 218,301 | 252.7 | 0.00864 | 0.109 |
| Canada | 6,116 | 26.7 | 0.00229 | 0.031 |
| Australia | 3,238 | 16.8 | 0.00193 | 0.026 |
| New Zealand | 323 | 3.4 | 0.00095 | 0.013 |
| United Kingdom | 5,451 | 57.3 | 0.00095 | 0.013 |
| Ireland | 142 | 3.6 | 0.00039 | 0.005 |

Slide 52 Stat 13, UCLA, Jon Dineen

Statistical independence

- Events A and B are *statistically independent* if knowing whether B has occurred gives no new information about the chances of A occurring, i.e. if $\text{pr}(A | B) = \text{pr}(A)$
- Similarly, $\text{P}(B | A) = \text{P}(B)$, since $\text{P}(B|A)=\text{P}(B \ \& \ A)/\text{P}(A) = \text{P}(A \ \& \ B)/\text{P}(A) = \text{P}(B)$
- If A and B are *statistically independent*, then

$$\text{pr}(A \text{ and } B) = \text{pr}(A) \times \text{pr}(B)$$

Slide 53 Stat 13, UCLA, Jon Dineen

People vs. Collins

TABLE 4.7.2 Frequencies Assumed by the Prosecution

| | | | |
|--------------------|----------------|---------------------------|------------------|
| Yellow car | $\frac{1}{10}$ | Girl with blond hair | $\frac{1}{3}$ |
| Man with mustache | $\frac{1}{4}$ | Black man with beard | $\frac{1}{10}$ |
| Girl with ponytail | $\frac{1}{10}$ | Interracial couple in car | $\frac{1}{1000}$ |

- The first occasion where a conviction was made in an American court of law, largely on statistical evidence, 1964. A woman was mugged and the offender was described as a wearing **dark cloths**, with **blond hair** in a **pony tail** who got into a **yellow car** driven by a **black male** accomplice with **mustache** and **beard**. The suspect brought to trial were picked out in a line-up and fit all of the descriptions. Using the *product rule for probabilities* an expert witness computed the chance that a random couple meets these characteristics, as 1:12,000,000.

Slide 55 Stat 13, UCLA, Jon Dineen

Formula summary cont.

- $\text{pr}(S) = 1$
- $\text{pr}(\bar{A}) = 1 - \text{pr}(A)$
- If A and B are mutually exclusive events, then $\text{pr}(A \text{ or } B) = \text{pr}(A) + \text{pr}(B)$
(here "or" is used in the inclusive sense)
- If A_1, A_2, \dots, A_k are mutually exclusive events, then $\text{pr}(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_k) = \text{pr}(A_1) + \text{pr}(A_2) + \dots + \text{pr}(A_k)$

Slide 56 Stat 13, UCLA, Jon Dineen

Formula summary cont.

Conditional probability

- Definition:

$$\text{pr}(A | B) = \frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)}$$
- Multiplication formula:

$$\text{pr}(A \text{ and } B) = \text{pr}(B|A)\text{pr}(A) = \text{pr}(A|B)\text{pr}(B)$$

Slide 57 Stat 13, UCLA, Jon Dineen

Formula summary cont.

Multiplication Rule under independence:

- If A and B are independent events, then

$$\text{pr}(A \text{ and } B) = \text{pr}(A) \text{pr}(B)$$
- If A_1, A_2, \dots, A_n are mutually independent,

$$\text{pr}(A_1 \text{ and } A_2 \text{ and } \dots \text{ and } A_n) = \text{pr}(A_1) \text{pr}(A_2) \dots \text{pr}(A_n)$$

Slide 58 Stat 13, UCLA, Jon Dineen

Law of Total Probability

- If $\{A_1, A_2, \dots, A_n\}$ are a partition of the sample space (mutually exclusive and $\cup A_i = S$) then for any event B

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_n)P(A_n)$$

Ex:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$

Slide 59 Stat 13, UCLA, Jon Dineen

Bayesian Rule

- If $\{A_1, A_2, \dots, A_n\}$ are a non-trivial partition of the sample space (mutually exclusive and $\cup A_i = S, P(A_i) > 0$) then for any non-trivial event and $B (P(B) > 0)$

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i) \times P(A_i)}{P(B)}$$

$$= \frac{P(A_i | B) \times P(A_i)}{\sum_{k=1}^n P(B | A_k) P(A_k)}$$

Slide 60 Stat 13, UCLA, Jon Dineen

Bayesian Rule

$$P(A_i) = \frac{P(A_i | B) \times P(A_i)}{\sum_{k=1}^n P(B | A_k) P(A_k)}$$

D = the test person has the disease.
 T = the test result is positive.

Ex: (Laboratory blood test) **Assume:** $P(\text{Disease}) = 0.005$ **Find:** $P(\text{Disease} | \text{positive Test}) = ?$

$P(\text{positive Test} | \text{no Disease}) = 0.01$ $P(D | T) = ?$

$$P(D | T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T | D) \times P(D)}{P(T | D) \times P(D) + P(T | D^c) \times P(D^c)}$$

$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = \frac{0.00475}{0.02465} = 0.193$$

Slide 61 Stat 13, UCLA, Jon Dineen

Classes vs. Evidence Conditioning

- Classes:** healthy (NC), cancer
- Evidence:** positive mammogram (pos), negative mammogram (neg)
- If a woman has a positive mammogram result, what is the probability that she has breast cancer?

$$P(\text{cancer} | \text{pos}) = \frac{P(\text{pos} | \text{cancer}) \times P(\text{cancer})}{P(\text{pos} | \text{cancer}) \times P(\text{cancer}) + P(\text{pos} | \text{healthy}) \times P(\text{healthy})}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} = ?$$

Slide 62 Stat 13, UCLA, Jon Dineen

Bayesian Rule (different data/example!)

| Test Results | True Disease State | | Total |
|--------------|----------------------------|-----------------------------|--------|
| | No Disease | Disease | |
| Negative | OK (0.98505) | False Negative II (0.00025) | 0.9853 |
| Positive | False Positive I (0.00995) | OK (0.00475) | 0.0147 |
| Total | 0.995 | 0.005 | 1.0 |

$$P(T \cap D^c) = P(T | D^c) \times P(D^c) = 0.01 \times 0.995 = 0.00995$$

Power of Test = $1 - P(T^c | D) = 0.00025 / 0.005 = 0.95$

Sensitivity: $TP / (TP + FN) = 0.00475 / (0.00475 + 0.00025) = 0.95$

Specificity: $TN / (TN + FP) = 0.98505 / (0.98505 + 0.00995) = 0.99$

Slide 63 Stat 13, UCLA, Jon Dineen

Examples – Birthday Paradox

- **The Birthday Paradox:** In a random group of N people, what is the change that at least two people have the same birthday?
- E.x., if $N=23$, $P>0.5$. Main confusion arises from the fact that in real life we rarely meet people having the same birthday as us, and we meet more than 23 people.
- The reason for such high probability is that any of the 23 people can compare their birthday with any other one, not just you comparing your birthday to anybody else's.
- There are N -Choose-2 = $20 \cdot 19 / 2$ ways to select a pair or people. Assume there are 365 days in a year, $P(\text{one-particular-pair-same-B-day}) = 1/365$, and
- $P(\text{one-particular-pair-failure}) = 1 - 1/365 \sim 0.99726$.
- For $N=20$, 20 -Choose-2 = 190. $E = \{\text{No 2 people have the same birthday is the event all 190 pairs fail (have different birthdays)}\}$, then $P(E) = P(\text{failure})^{190} = 0.99726^{190} = 0.59$.
- Hence, $P(\text{at-least-one-success}) = 1 - 0.59 = 0.41$, quite high.
- Note: for $N=42 \rightarrow P > 0.9 \dots$

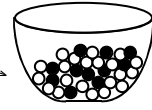
Slide 64 Stat 13, UCLA, Jon Dineen

The two-color urn model

N balls in an urn, of which there are

M black balls

$N - M$ white balls



Sample n balls and count $X = \#$ black balls in sample

We will compute the probability distribution of the R.V. X

Slide 65 Stat 13, UCLA, Jon Dineen

The biased-coin tossing model



toss 1
 $\text{pr}(H) = p$

toss 2
 $\text{pr}(H) = p$

toss n
 $\text{pr}(H) = p$

Perform n tosses and count $X = \#$ heads

We also want to compute the probability distribution of this R.V. X !
Are the two-color urn and the biased-coin models related? How do we present the models in mathematical terms?

Slide 66 Stat 13, UCLA, Jon Dineen

The answer is: Binomial distribution

- The distribution of the number of heads in n tosses of a biased coin is called the **Binomial distribution**.

Slide 67 Stat 13, UCLA, Jon Dineen

Binomial(N, p) – the probability distribution of the number of Heads in an N -toss coin experiment, where the probability for Head occurring in each trial is p .
E.g., Binomial(6, 0.7)

| | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|------------|-----------------------|-------|-------|-------|-------|-------|-------|-------|
| Individual | $\text{pr}(X = x)$ | 0.001 | 0.010 | 0.060 | 0.185 | 0.324 | 0.303 | 0.118 |
| Cumulative | $\text{pr}(X \leq x)$ | 0.001 | 0.011 | 0.070 | 0.256 | 0.580 | 0.882 | 1.000 |

For example $P(X=0) = P(\text{all 6 tosses are Tails}) =$

$$(1 - 0.7)^6 = 0.3^6 = 0.001$$

Slide 68 Stat 13, UCLA, Jon Dineen

Binary random process

The *biased-coin tossing model* is a physical model for situations which can be characterized as a series of trials where:

- each trial has only **two outcomes**: *success* or *failure*;
- $p = P(\text{success})$ is the same for every trial; and
- trials are **independent**.

- The distribution of $X =$ number of successes (heads) in N such trials is

Binomial(N, p)

Slide 69 Stat 13, UCLA, Jon Dineen

Sampling from a finite population – Binomial Approximation

If we take a sample of size n

- from a much larger population (of size N)
- in which a proportion p have a characteristic of interest, then the distribution of X , the number in the sample with that characteristic,
- is approximately Binomial(n, p).
 - (Operating Rule: Approximation is adequate if $n/N < 0.1$.)
- Example, polling the US population to see what proportion is/has-been married.

Slide 70 Stat 13, UCLA, Jon Dinger

Binomial Probabilities – the moment we all have been waiting for!

- Suppose $X \sim \text{Binomial}(n, p)$, then the probability

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}, \quad 0 \leq x \leq n$$

- Where the binomial coefficients are defined by

$$\binom{n}{x} = \frac{n!}{(n-x)! x!}, \quad n! = 1 \times 2 \times 3 \times \dots \times (n-1) \times n$$

n-factorial

Slide 71 Stat 13, UCLA, Jon Dinger

Expected values

- The game of chance: cost to play: \$1.50; Prices {\$1, \$2, \$3}, probabilities of winning each price are {0.6, 0.3, 0.1}, respectively.
- Should we play the game? What are our chances of winning/loosing?

| Prize (\$) | x | 1 | 2 | 3 | |
|---------------------------------------|-------|-------------------------------|---------------|---------------|----------------|
| Probability | pr(x) | 0.6 | 0.3 | 0.1 | |
| What we would "expect" from 100 games | | | | | |
| Number of games won | | 0.6 × 100 | 0.3 × 100 | 0.1 × 100 | add across row |
| \$ won | | 1 × 0.6 × 100 | 2 × 0.3 × 100 | 3 × 0.1 × 100 | Sum |
| Total prize money = Sum; | | Average prize money = Sum/100 | | | |
| | | = 1 × 0.6 + 2 × 0.3 + 3 × 0.1 | | | |
| | | = 1.5 | | | |

Theoretically Fair Game: price to play EQ the expected return!

Slide 72 Stat 13, UCLA, Jon Dinger

Definition of the expected value, in general.

- The expected value:

$$E(X) = \sum_{\text{all } x} x P(x) \left(= \int x P(x) dx \right)$$

- = Sum of (value times probability of value)

Slide 73 Stat 13, UCLA, Jon Dinger

Example

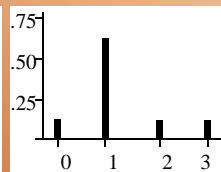
In the at least one of each or at most 3 children example, where $X = \{\text{number of Girls}\}$ we have:

| X | 0 | 1 | 2 | 3 |
|-------|---------------|---------------|---------------|---------------|
| pr(x) | $\frac{1}{8}$ | $\frac{5}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ |

$$E(X) = \sum_x x P(x)$$

$$= 0 \times \frac{1}{8} + 1 \times \frac{5}{8} + 2 \times \frac{1}{8} + 3 \times \frac{1}{8}$$

$$= 1.25$$



Slide 74 Stat 13, UCLA, Jon Dinger

The expected value and population mean

$\mu_x = E(X)$ is called the *mean* of the distribution of X .

$\mu_x = E(X)$ is usually called the *population mean*.

μ_x is the point where the bar graph of $P(X = x)$ balances.

Slide 75 Stat 13, UCLA, Jon Dinger

Population standard deviation

The *population standard deviation* is
 $sd(X) = \sqrt{E[(X - \mu)^2]}$

Note that if X is a RV, then $(X - \mu)$ is also a RV, and so is $(X - \mu)^2$. Hence, the *expectation*, $E[(X - \mu)^2]$, makes sense.

Slide 76 Stat 13, UCLA, Jon Dineen

Population mean & standard deviation

Expected value: $E(X) = \sum_x xP(X = x)$

Variance $Var(X) = \sum_x (x - E(x))^2 P(X = x)$

Standard Deviation $SD(X) = \sqrt{Var(X)} = \sqrt{\sum_x (x - E(x))^2 P(X = x)}$

Slide 77 Stat 13, UCLA, Jon Dineen

For the Binomial distribution . . . Mean & SD

$$E(X) = np,$$

$$sd(X) = \sqrt{np(1-p)}$$

Slide 78 Stat 13, UCLA, Jon Dineen

The Normal Distribution

- Recall: in chapter 2 we used histograms to represent frequency distributions.
 - We can think of a histogram as an approximation of the true population distribution.
- A smooth curve representing a frequency distribution is called a **density curve**.

Slide 79 Stat 13, UCLA, Jon Dineen

Linear Scaling (affine transformations) $aX + b$

Why is that so?

$$E(aX + b) = a E(X) + b \quad SD(aX + b) = |a| SD(X)$$

$$E(aX + b) = \sum_{x=0}^n (a x + b) P(X = x) =$$

$$\sum_{x=0}^n a x P(X = x) + \sum_{x=0}^n b P(X = x) =$$

$$a \sum_{x=0}^n x P(X = x) + b \sum_{x=0}^n P(X = x) =$$

$$aE(X) + b \times 1 = aE(X) + b.$$

Slide 80 Stat 13, UCLA, Jon Dineen

Linear Scaling (affine transformations) $aX + b$

And why do we care?

$$E(aX + b) = a E(X) + b \quad SD(aX + b) = |a| SD(X)$$

-E.g., say the rules for the game of chance we saw before change and the new pay-off is as follows: $\{\$0, \$1.50, \$3\}$, with probabilities of $\{0.6, 0.3, 0.1\}$, as before. What is the newly expected return of the game? Remember the old expectation was equal to the entrance fee of \$1.50, and the game was fair!

$$Y = 3(X-1)/2$$

$$\{\$1, \$2, \$3\} \rightarrow \{\$0, \$1.50, \$3\},$$

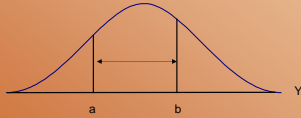
$$E(Y) = 3/2 E(X) - 3/2 = 3 / 4 = \$0.75$$

And the game became clearly biased. Note how easy it is to compute $E(Y)$.

Slide 81 Stat 13, UCLA, Jon Dineen

The Normal Distribution

- The normal distribution is a bell shaped, symmetric density curve

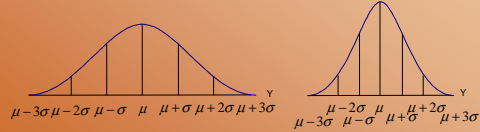


- Area under density curve between a and b is equal to the proportion of Y values between a and b.
- The area under the whole curve is equal 1.0

Slide 82 Stat 13, UCLA, Jon Dineen

The Normal Distribution

- Each normal curve is characterized by its μ and σ



- If random variable Y is normal with mean μ and standard deviation σ , we write

$$Y \sim N(\mu, \sigma^2)$$

Slide 83 Stat 13, UCLA, Jon Dineen

The Normal Distribution

- A normal density curve can be summarized with the following formula:

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

- Every normal curve uses this formula, what makes them different is what gets plugged in for μ and σ
- Each normal curve is centered at μ and the width depends on σ
- (small = tall, large = short/wide).

Slide 84 Stat 13, UCLA, Jon Dineen

Areas under the normal curve

- Because each normal curve is the result of a single formula the areas under the normal curve have been computed and tabulated for ease of use.

- The Standard Scale

- Any normal curve can be converted into a normal curve with
- $\mu = 0$ and $\sigma = 1$. This is called the standard normal.

Slide 85 Stat 13, UCLA, Jon Dineen

Areas under the normal curve

- The process of converting normal data to the standard scale is called standardizing.
- To convert Y into Z (a z-score) use the following formula:

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

- What does a z-score measure?

Slide 86 Stat 13, UCLA, Jon Dineen

Areas under the normal curve

- Table 3 (also in front of book) gives areas under the standard normal curve

Example: Find the area that corresponds to $z < 2.0$

- Always good to draw a picture!

Example: Find the area that corresponds to $z > 2.0$

Example: Find the area that corresponds to $1.0 < z < 2.0$

Example: Find the area that corresponds to $z < 2.56$

Tables are antiquated → Use tools like SOCR (socr.ucla.edu)

Slide 87 Stat 13, UCLA, Jon Dineen

Relationship to the Empirical Rule

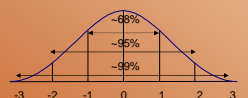
- $\bar{y} \pm s \approx 68\%$
- Recall the Empirical Rule $\bar{y} \pm 2s \approx 95\%$
- $\bar{y} \pm 3s \approx 99\%$
- How can we use the standard normal distribution to verify the properties of the empirical rule?

The area: $-1 < z < 1 = 0.8413 - 0.1587 = 0.6826$
 The area: $-2.0 < z < 2.0 = 0.9772 - 0.0228 = 0.9544$
 The area: $-3.0 < z < 3.0 = 0.9987 - 0.0013 = 0.9974$

Slide 88 Stat 13, UCLA, Jon Dineen

Relationship to the Empirical Rule

- Visually:
- http://socr.stat.ucla.edu/htmls/SOCR_Distributions.html



Slide 89 Stat 13, UCLA, Jon Dineen

Application to Data

Example: Suppose that the average systolic blood pressure (SBP) for a Los Angeles freeway commuter follows a normal distribution with mean 130 mmHg and standard deviation 20 mmHg.

Find the percentage of LA freeway commuters that have a SBP less than 100.

- First step: Rewrite with notation!
 $Y \sim N(130, 20)$

Slide 90 Stat 13, UCLA, Jon Dineen

Application to Data

- Second step: Identify what we are trying to solve!
 Find the percentage for: $y < 100$
- Third step: Standardize

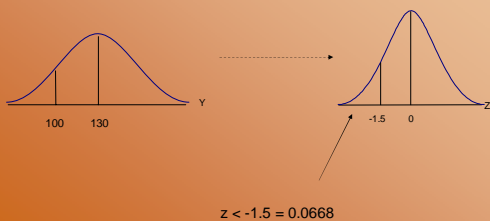
$$Z = \frac{Y - \mu}{\sigma} = \frac{100 - 130}{20} = -1.5$$
- Fourth Step: Use the standard normal table to solve
 $y < 100 = z < -1.5 = 0.0668$

Therefore approximately 6.7% of LA freeway commuters have SBP less than 100 mmHg.

Slide 91 Stat 13, UCLA, Jon Dineen

Application to Data

- Visually



Slide 92 Stat 13, UCLA, Jon Dineen

Application to Data

- Try these:
 - What percentage of LA freeway commuters have SBP greater than 155 mmHg?
 - Between 120 and 175?
- Can also be interpreted in terms of probability.
 - What is the probability that a randomly selected freeway commuter will have a SBP less than 100?

$$P(Y < 100) = 0.0668$$

Slide 93 Stat 13, UCLA, Jon Dineen

Normal approximation to Binomial

- Suppose $Y \sim \text{Binomial}(n, p)$
- Then $Y = Y_1 + Y_2 + Y_3 + \dots + Y_n$, where
 - $Y_k \sim \text{Bernoulli}(p)$, $E(Y_k) = p$ & $\text{Var}(Y_k) = p(1-p) \rightarrow$
 - $E(Y) = np$ & $\text{Var}(Y) = np(1-p)$, $\text{SD}(Y) = (np(1-p))^{1/2}$
 - **Standardize Y :**
 - $Z = (Y - np) / (np(1-p))^{1/2}$
 - By CLT $\rightarrow Z \sim N(0, 1)$. So, $Y \sim N(np, (np(1-p))^{1/2})$
- **Normal Approx to Binomial is reasonable when $np \geq 10$ & $n(1-p) \geq 10$** (p & $(1-p)$ are NOT too small relative to n).

Slide 94 Stat 13, UCLA, Jon Dineen

Normal approximation to Binomial – Example

- **Roulette wheel investigation:**
- Compute $P(Y \geq 58)$, where $Y \sim \text{Binomial}(100, 0.47)$ –
 - The proportion of the Binomial(100, 0.47) population having more than 58 reds (successes) out of 100 roulette spins (trials).
 - Since $np = 47 \geq 10$ & $n(1-p) = 53 \geq 10$ **Normal approx is justified.**
- $Z = (Y - np) / \text{Sqrt}(np(1-p)) = 100 * 0.47 / \text{Sqrt}(100 * 0.47 * 0.53) = 2.2$
- $P(Y \geq 58) \leftrightarrow P(Z \geq 2.2) = 0.0139$
- True $P(Y \geq 58) = 0.177$, using SOCR (demo!)
- Binomial approx useful when no access to SOCR avail.

Roulette has 38 slots
 18 red 18 black 2 neutral

Slide 95 Stat 13, UCLA, Jon Dineen

Percentiles

- Divides the distribution into 100 equal parts.
 - The p^{th} percentile means $p\%$ of the observations lie below and $1-p\%$ above
- Example: Suppose we want to find the value z that cuts off the top 2.5% of the distribution.

$P(Z > Z_{0.025}) = 0.025$

Your author calls this Z_α

What is $Z_{0.025}$ (ie. What Z is the cut point for the 97.5th percentile)?

Slide 96 Stat 13, UCLA, Jon Dineen

Percentiles

- This is the reverse situation from before
 - Now we have what is inside the curve and we need the z-score.

We also know that $P(Z < Z_{0.025}) = 0.975$

Using the table we can solve: $Z_{0.025} = 1.96$

Slide 97 Stat 13, UCLA, Jon Dineen

Percentiles

Example (con't): From the LA freeway commuters, find the SBP that is the 90th percentile.

$P(Z < Z_{0.10}) = 0.90$

Need to choose the z-score that will give the area closest to 0.90!

Using the table we can solve: $Z_{0.10} = 1.28$

Slide 98 Stat 13, UCLA, Jon Dineen

Percentiles

Example (con't): We're half way there...

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

$$1.28 = \frac{Y^* - 130}{20}$$

$$Y^* = 155.6 \text{ mmHg}$$

Slide 99 Stat 13, UCLA, Jon Dineen

Assessing Normality

- How can we tell if our data is normally distributed?
- Several methods for checking normality
 - Mean = Median
 - Empirical Rule
 - Check the percent of data that within 1 sd, 2 sd and 3 sd (should be approximately 68%, 95% and 99.7%).
 - Histogram or dotplot
 - Normal Probability Plot
- Why do we care if the data is normally distributed?

Slide 100 Stat 13, UCLA, Jon Dineen

Normal Probability Plots

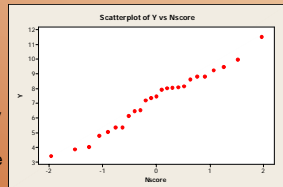
- A normal probability plot is a graph that is used to assess normality in a data set.
- When we look at a normal plot we want to see a straight line.
 - This means that the distribution is approximately normal.
 - Sometimes easier to see if a line is straight, than if a histogram is bell shaped.

Slide 101 Stat 13, UCLA, Jon Dineen

Normal Probability Plots

- This is how the plot works:

- We take the data and plot it against normal scores
- To compute normal scores we take expected values of ordered observations from a sample of size n that is normally distributed $N(0,1)$.
- When we then compare these "normal scores" to the actual y values on a graph, if the data were normal, we will see our straight line.



Slide 102 Stat 13, UCLA, Jon Dineen

Normal Probability Plots

Example: height example from book p.134-135

Suppose we have the height for 11 women.

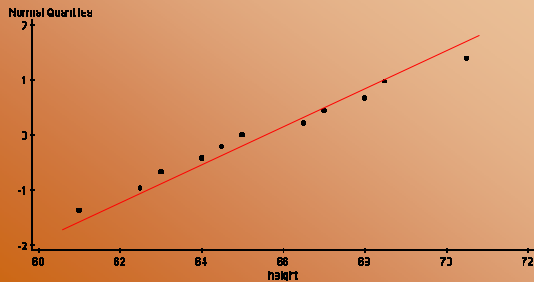
| height (in) | Nscore |
|-------------|----------|
| 61.0 | -1.59322 |
| 62.5 | -1.06056 |
| 63.0 | -0.72791 |
| 64.0 | -0.46149 |
| 64.5 | -0.22469 |
| 65.0 | 0.00000 |
| 66.5 | 0.22469 |
| 67.0 | 0.46149 |
| 68.0 | 0.72791 |
| 68.5 | 1.06056 |
| 70.5 | 1.59322 |

Calculated using SOCR, slightly different than formula from text.

Slide 103 Stat 13, UCLA, Jon Dineen

Normal Probability Plots

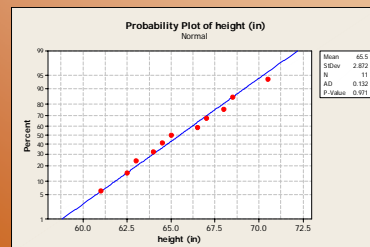
Example (cont'): Normal probability plot



Slide 104 Stat 13, UCLA, Jon Dineen

Normal Probability Plots

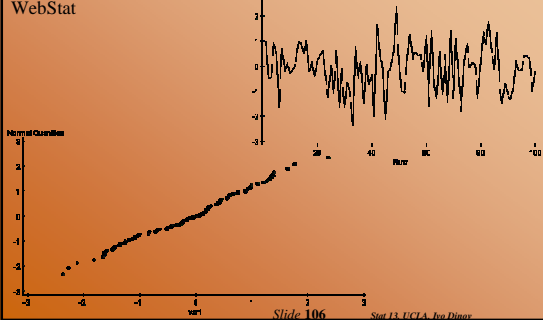
Example (cont'): The normal plot we will use:



Slide 105 Stat 13, UCLA, Jon Dineen

Normal Probability Plots - Simulation

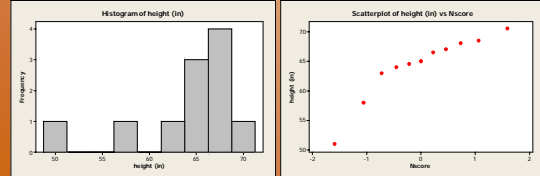
Example: Random Sampling from Normal (0,5): Raw Sample + QQPlot
http://socr.stat.ucla.edu/htmls/SOCR_Modeler.html, Data Generation + WebStat



Slide 106 Stat 13, UCLA, Jon Dineen

Diagnostics

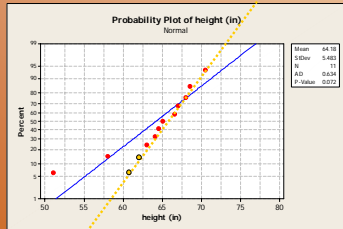
- If the bottom of the distribution bends down, this implies that the y values at the lower end are small.
 - In other words skewed left



Slide 107 Stat 13, UCLA, Jon Dineen

Diagnostics

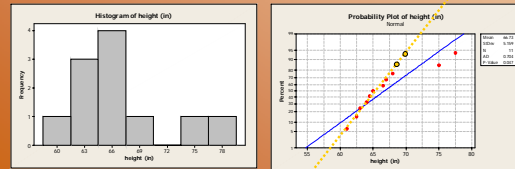
- On the normal probability plot this would look like:



Slide 108 Stat 13, UCLA, Jon Dineen

Diagnostics

- This next plot would also not be considered normally distributed. What do they tell you about the shape of the distribution? Why?



Slide 109 Stat 13, UCLA, Jon Dineen