1.1 Introduction to Probability

- **Basis/Genesis: Games of chance**
- **Applications:**
  - **Modeling genetic mutations**
  - **Modeling queues in banks, computer operating systems, networks, etc.**
  - **Modeling atmospheric turbulence and weather forecasting**
  - **Computer simulation**
  - **Reliability (etc., see p.1 in text)**
- **Basis for *statistical inference***

1.2 Sample Spaces
(A Quick Review of Set Theory, Venn Diagrams)

***Experiment*: An activity that generates random outcomes**

***Sample Space*: Set (Ω) of possible outcomes from an experiment. An element of Ω is called ω.**

Ω finite: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$
Ω infinite (but countable): $\Omega = \{\omega_1, \omega_2, \ldots \}$
Ω infinite (uncountable): $\Omega = \{\omega | 0 \leq \omega_2, \leq 1.5 \}$

Examples of Sample Spaces

Example 1.2-1 A consumer rates a product on a 1-5 (Likert) scale: 1 = poor, 2 = fair, 3 = expected, 4 = good, 5 = excellent.

Ω =

Example 1.2-2 Count the number of customers who arrive between 9:00AM and 12:00AM at a store.

Ω =

Example 1.2-3 Let t denote the length of time between successive earthquakes in Marin County, CA.

Ω =

Events

***Event*: An event, A, is any subset of Ω. (We write $A \subset \Omega$ for "A is a subset of Ω" or $\Omega \supset A$ for "Ω contains A.")**

Example 1.2-4 A $\equiv$ event that a consumer rates a product "better than expected" using the above 5-point scale.

A =

Events

Example 1.2-5 Write out the following events as sets.

A ≡ event that the number of customers in the store between 9:00AM and 12:00AM is less than 10.

B ≡ event that this number exceeds 5.

D ≡ event that this number exceeds 9.

E ≡ event that this number is an even number.

Unions of Events

A *union* of two events C = A∪ B = A "union" B = A "or" B is the event that either A occurs or B occurs or both occur.

Example 1.2-6 (See example 1.2-5.) Let C = A∪ E. Find C.

Example 1.2-7 (See example 1.2-5) Let C = A∪ B. Find C

Intersections and Compliments of Events

An *intersection* of two events $C = A \cap B$ = A "intersect" B = A "and" B is the event that both A and B occur.

Example 1.2-8 (See example 1.2-5.) Let $C = A \cap B$. Find C.

A *compliment* of an event A, denoted A^c, is the event that A does not occur

Example 1.2-9 (See example 1.2-5). Find A^c

Disjoint (Mutually Exclusive) Sets

Two sets A and B are said to be disjoint if A and B have no events in common: $A \cap B = \{\ \} \equiv \varnothing$.

Example 1.2-10 (See Example 1.2-5.) Find $A \cap D$.

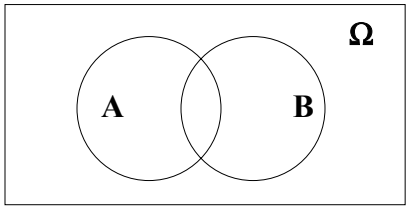
Laws of Set Theory

Commutative: $A \cup B = B \cup A$
$A \cap B = B \cap A$

Associative: $(A \cup B) \cup C = A \cup (B \cup C)$
$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Venn Diagrams



$A \cup B$:

$A \cap B$

1.3 Probability Measures

A *probability measure* on Ω is a function P from the subsets of Ω to the real numbers that satisfies the following axioms:

1. $P(\Omega) = 1$
2. If $A \subset \Omega$, then $P(A) \geq 0$
3. If A_1 and A_2 are disjoint, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Probability Measures: Properties

Property A: $P(A^c) = 1 - P(A)$

Justification:

From Axiom 3, $P(A \cup A^c) = P(\Omega) = P(A) + P(A^c)$.

From Axiom 1, $P(\Omega) = 1$.

Therefore: $1 = P(A) + P(A^c)$.

Property B: $P(\varnothing) = 0$.

Justification: Let $A = \varnothing$; then $A^c = \Omega$

From Property A: $P(\varnothing) = 1 - P(\Omega) = 0$.

Probability Measures: Properties

Property C: If $A \subset B$, then $P(A) \leq P(B)$.

Justification:

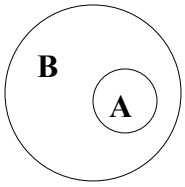
$B = A \cup (A^c \cap B)$,

(a union of disjoint sets), so

$P(B) = P(A) + P(A^c \cap B)$

$\geq P(A)$

(since $P(A^c \cap B) \geq 0$).



Probability Measures: Properties

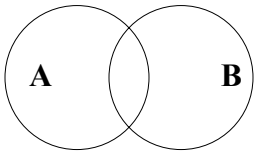
Property D (*Addition Law*):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Intuitive justification:

Using $P(A \cup B) = P(A) + P(B)$

"double counts" $P(A \cap B)$.



To fix this error, subtract $P(A \cap B)$ once to get:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Coin Tossing

Example 1.3-1 Flip a fair coin twice. H_1 = head on first toss, H_2 = head on second toss, T_1 and T_2 denote tails on first and second tosses.

Ω =

If the coin is balanced, we can assume all four elementary events (ω's) are equally likely:

$P(H_1\ H_2) = P(H_1\ T_2) = P(T_1\ H_2) = P(T_1\ T_2) =$

$C \equiv H_1 \cup H_2$. Find P(C).

Statistical Risks of AIDS Transmission

Example 1.3-2 From the LA Times, 8/24/87:

> **Several studies of sexual partners of people infected with the virus show that a single act of unprotected vaginal intercourse has a surprisingly low risk of infecting the uninfected partner--perhaps one in 100 to one in 1000. For an average, consider the risk to be one in 500. If there are 100 acts of intercourse with an infected partner, the odds of infection increase to one in five.**
>
> **Statistically, 500 acts of intercourse with one infected partner or 100 acts with five partners lead to a 100% probability of infection (statistically, not necessarily in reality).**

What's wrong with this analysis? Find the probability of transmission in two acts, where T_1 denotes tranmission on first act, and T_2 denotes transmission on the second.

1.4 Computing Probabilities: Counting Methods

Classical Definition of Probability: If Ω is finite such that $\Omega = \{\omega_1, \ldots, \omega_N\}$ *and all elements of Ω have equal probability*, then the probability of any event A is:

$$P(A) = \frac{\text{Number of ways A can occur}}{\text{N, the number of elements in } \Omega}$$

Example 1.4-1: Choosing the best medical treatment. Estimate P(S) for treatments 1 and 2 for males.

	For Males Only			
	Success (S)	Failure (F)	Total (N)	P(S)
Treatment 1	60	20	80	
Treatment 2	100	50	150	

Example 1.4-1 (Continued)

Determine P(S) for treatments 1 and 2 for females:

	For Females Only			
	Success (S)	Failure (F)	Total (N)	P(S)
Treatment 1	40	80	120	
Treatment 2	10	30	40	

Now combine the tables and determine P(S) for treatments 1 and 2 for human beings (male or female):

	For Males and Females			
	Success (S)	Failure (F)	Total (N)	P(S)
Treatment 1	100	100	200	
Treatment 2	110	80	190	

(This is an example of *Simpson's Paradox*. Conclusion: not always safe to collapse tables into smaller table)

1.4.1 The Multiplication Principle

If experiment one has n_1 outcomes, experiment two has n_2 outcomes, …, and the pth experiment has n_p outcomes, then for the p experiments the total number of outcomes is:

$$\prod_{i=1}^{p} n_i = n_1 \times n_2 \times \cdots \times n_p$$

Example 1.4-2 Playing cards have 13 face values and 4 suits. How many combinations of face values and suits are there?

Example 1.4-3 How many 8-bit binary words are possible?

1.4.2 Permutations and Combinations

***Permutation*: An *ordered* arrangement of objects.**

Suppose C = {c_1, c_2, …, c_n} and we choose r elements *without replacement* from C and list them in order. How many orderings (permutations) are possible?

$$\begin{aligned} {}_nP_r &= n \times (n-1) \times (n-2) \times \cdots \times (n-[r-1]) \\ &= n \times (n-1) \times \cdots \times (n-[r-1]) \times \frac{(n-r) \times (n-r-1) \times \cdots \times 1}{(n-r) \times (n-r-1) \times \cdots \times 1} \\ &= \end{aligned}$$

Example 1.4-4 Minnesota auto license plates have six characters: 3 letters followed by 3 numbers. How many plates are possible? (Be careful!)

Permutations: Examples

Example 1.4-5 How many plates are possible that contain no repeats of letters or numbers?

Example 1.4-6 If all sequences of six characters are equally likely, find the probability of the event A, that a license plate on a new car will have no repeats.

Birthday Problem

Example 1.4-6 A room contains n people. What is the probability that *at least* two of them have the same birthday? (Assume 365 birthdays are equally likely.)

$A \equiv$ event of one or more matches; $A^c \equiv$ event of no matches

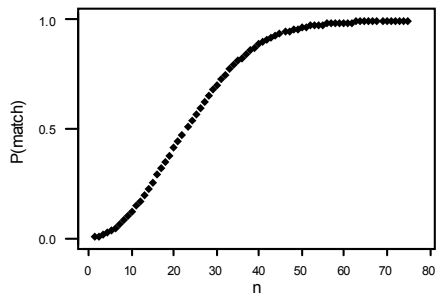
$$P(A) = 1 - P(A^c)$$

$$P(A) = 1 - \frac{\text{Number of orderings having } n \text{ birthdays with no match}}{\text{Number of orderings of } n \text{ birthdays}}$$

$$=$$

Birthday Problem: Calculations and Plot

n	P(match)
4	0.016
16	0.284
23	0.507
32	0.753
40	0.891
56	0.988



Combinations

Suppose I select three students from the class to work a problem together. Here, order is *not* important, so the combination {Mike, Joe, Heidi} is the same combination as {Heidi, Mike, Joe}.

Similarly the card hand {Ace, Ace, 2, 5, 9} is the same card hand as {2, Ace, 5, Ace, 9}. Order matters not.

Combinations: The number of ways to choose r objects from n (without regard to order) is:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Combinations

Relation between permutations and combinations: Any combination or r objects has ${}_rP_r$ = r! distinct orderings. Therefore:

$$\binom{n}{r} = {}_nP_r \div r!$$

Note 1: $\binom{n}{r}$ is called the *binomial coefficient* because:

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Check: Suppose n = 3:

Note 2: Let a = b = 1. Then:

Combinations

Example 1.4-7 Five cards are drawn at random from a deck of 52.

(a) How many 5-card hands are there?

(b) How many 5-card hands are there that contain four aces?

(c) Find P(4 aces) in draw of five cards.

Combinations

Example 1.4-10: Quality control acceptance sampling.

You receive product in lots of size n. Suppose k of these n items are defective, and n-k are not defective. Choose r items at random for inspection.

(a) How many samples are possible?

(b) How many of these contain exactly m defectives?

Example 1.4-7: The Capture-Recapture Method

Used to estimate the size n of a wildlife population.

Phase 1: Capture *t* animals, tag, and return to the population.

Phase 2: Capture *m* animals, count the number that were captured previously. Call this number *r* for "recaptured."

Suppose t = 10, m =20 and r = 4. Give a formula for P(*r*)

Capture-Recapture (continued)

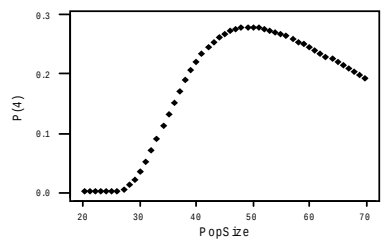
So how does this help us determine *n*?

We can *estimate n* using the *method of maximum likelihood:* the "best" value of n will maximize the probability P(4) of what we observed (4).

Calculations

n	P(n)
45	0.269
46	0.274
47	0.277
48	0.279
49	0.280
50	0.280
51	0.279
52	0.278
53	0.275

Plot P(4) vs *n*:



Result: "Most likely" value (MLE) of n is:

The Multinomial Coefficient

The number of ways that n objects can be grouped into r classes with n_i in the ith class, i = 1,…,r and $\Sigma n_i = n$ is:

$$\binom{n}{n_1\ n_2 \cdots n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

Justification: From the binomial coefficient and the multiplication principle:

$$\binom{n}{n_1\ n_2 \cdots n_r} = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r}$$

Cancellation yields the result above.

Multinomial Coefficient

Example 1.4-8 In how many ways could I group our class of 17 students into three groups of size four and one group of size 5 for group project work?

Example 1.4-9 How many ways can I pair my team of 10 tennis players together for practice matches?

Note on the Multimomial Coefficient

The multimomial coefficients occur in the expansion:

$$(x_1 + \cdots + x_r)^n = \sum \binom{n}{n_1\ n_2 \cdots n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

where the sum is over all nonnegative $n_1, \ldots, n_r$ such that $\Sigma n_i = n$.

1.5 Conditional Probability

Example 1.5-1: Personality Type and Exercise

3182 people without cardiovascular disease were cross-classified by two factors: personality type (A or B) and exercise regimen:

	Personality Type		
Exercise Regimen	**A**	**B**	**Totals**
Exercise regularly (E+)	483	477	960
Do not exercise regularly (E-)	1101	1121	2222
Totals	1584	1598	3182

Assuming this sample is representative (or that this finite group is a population) we can list these frequencies as probabilities relative to 3182. For example:

$P(E+\cap B) =$

Example 1.5.1 (Continued)

Converting the table:

	Personality Type		
Exercise Regimen	**A**	**B**	**Totals**
Exercise regularly (E+)	0.152	0.150	0.302
Do not exercise regularly (E-)	0.346	0.352	0.698
Totals	0.498	0.502	1.000

Notes:

$P(A) =$ **$P(E+\cap A) =$**

$P(B) =$ **$P(E+\cap B) =$**

$P(E+) =$ **$P(E-\cap A) =$**

$P(E-) =$ **$P(E-\cap B) =$**

Example 1.5.1 (Continued)

We might ask: What is the probability that a person exercises regularly, if we restrict attention only to type-A individuals? In symbols:

P(E+|A) = **"the probability of E+ occurring *given* A has occurred**

= **"the probability of E+ given A**

=

Equivalently:

$$P(E+|A) = \frac{P(E+ \cap A)}{P(A)}$$

Conditional Probability and Multiplication Law

Definition: If $P(B) \neq 0$, the *conditional probability* of A given B is:

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

Multiplying through on both sides by P(B) yields:

The Multiplication Law: If $P(B) \neq 0$, then:

$$P(A \cap B) = P(A|B)P(B)$$

Conditional Probability Examples

Example 1.5-2 An urn contains 3 red balls and one blue ball. Balls are selected without replacement. Let R_1 and R_2 denote the events that red balls are selected on the first and second draws, respectively. Find $P(R_1 \cap R_2)$.

Example 1.5-3 In a draw of two cards from a 52-card deck, find the probability of drawing two aces.

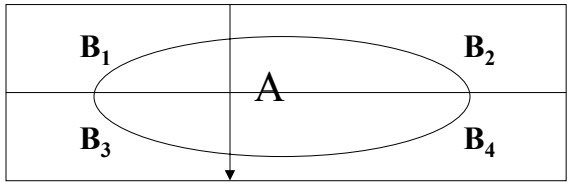
Example 1.5-4 Let $A \equiv$ event that it is raining, and let B denote the event that it is cloudy. If $P(B) = .3$ and $P(A|B) = .2$, find $P(A \cap B)$.

Law of Total Probability

Let $B_1, B_2, \ldots, B_n$ be such that $\cup B_i = \Omega$, and $B_i \cap B_j = \varnothing$ for $i \neq j$, with $P(B_i) > 0$ for all i. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Proof: (see text). Intuition is easy from picture:



Law of Total Probability (continued)

So:

$$A = (A \cap B_1) \cup (B \cap B_2) \cup (A \cap B_3) \cup (A \cap B_4)$$

And:

$$P(A) = \sum_{i=1}^{4} P(A \cap B_i) \quad \text{(by disjointness of } \{B_i\}\text{)}$$

$$= \sum_{i=1}^{4} P(A \mid B_i)P(B_i) \quad \text{(by Multiplication Law)}$$

Example 1.5-5 (See Example 1.5-2). Find $P(R_2)$

$\Omega = R_1 \cup B_1$ where R_1 and B_1 are disjoint. Hence:

$P(R_2) =$

Example 1.5-6 Occupational Mobility in England

Matrix of Transition Probabilities				
	Son's Occupational Status			
Father's Occupational Status	Upper (U_2)	Middle (M_2)	Lower (L_2)	Total
Upper (U_1)	0.45	0.48	0.07	1.00
Middle (M_1)	0.05	0.70	0.25	1.00
Lower (L_1)	0.01	0.50	0.49	1.00

So, for example:

P(Rich son|Rich father) = $P(U_2|U_1) = .45$

P(Poor son|Rich father) = $P(L_2|U_1) = .07$

Suppose that in the father's generation, 10% were upper-level, 40% were middle level, and 50% were lower-level. What is the percentage of men in the son's generation who are middle-level--i.e., what is $P(M_2)$?

Example 1.5-6 Occupational Mobility (continued)

Matrix of Transition Probabilities				
	Son's Occupational Status			
Father's Occupational Status	Upper (U_2)	Middle (M_2)	Lower (L_2)	Total
Upper (U_1)	0.45	0.48	0.07	1.00
Middle (M_1)	0.05	0.70	0.25	1.00
Lower (L_1)	0.01	0.50	0.49	1.00

U_1, M_1, and L_1 are disjoint, and $U_1 \cup M_1 \cup L_1 = \Omega$. So by the law of total probability:

The Inverse Problem and Bayes Rule

Suppose we wanted to solve the following: if a son has occupational level U_2, what is the probability that the father's status was U_1?

Example of an "inverse" problem. Given an "effect," want to find the probability of a particular "cause."

$P(U_1|U_2) =$

Bayes Rule

Let A and $B_1, \ldots, B_n$ be events where the B_i are disjoint, $\bigcup_{i=1}^{n} B_i = \Omega$ and $P(B_i) > 0$ for all i. Then:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Example 1.5-7 Polygraph (Lie-Detector) Tests

+ ≡ event that polygraph test is positive--says subject is lying

- ≡ event that polygraph test is negative

T ≡ event that subject is actually telling the truth

L ≡ event that subject is actually lying

From recent studies:

P(+|L) = .88 **P(-|L) = .12**

P(+|T) = .14 **P(-|T) = .86**

Example 1.5-7 Polygraph (Lie-Detector) Tests

Case 1: Suppose in the general population, 99% are truthful: P(T) = .99 and P(L) = .01. These are the *prior probabilities*. Find P(T|+), the *posterior probability* of telling the truth given the polygraph indicated a lie.

Example 1.5-7 Polygraph (Lie-Detector) Tests

Case 2: Suppose now that the general population is untrustworthy (say prison inmates or lawyers), and only 30% are truthful: P(T) = .30 and P(L) = .70. Now find the posterior probability P(T|+).

Note the effect of the prior probabilities on the posteriors!

1.6 Independence

Definition: Two events A and B are *independent* if knowing that one occurred gives us no new information about whether or not the other occurred. That is, A and B are independent if:

$P(A|B) = P(A)$ (or equivalently if $P(B|A) = P(B)$)

Note: If A and B are independent, the Multiplication Law simplifies to:

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

Independence: Examples

Example 1.6-1 (See Examples 1.5-2 and 1.5-5)

Urn contains 3 red balls and one blue ball; we draw two balls, one at a time without replacement. Are the events R_1 and R_2 independent? Recall $P(R_2) = .75$.

Example 1.6-2 An airplane has two engines that fail independently with probability $p = .001$. Find the probability that both engines fail.

Example 1.6-3: Infectivity of AIDS

Recall Example 1.3-2: From the LA Times, 8/24/87:

> **Several studies of sexual partners of people infected with the virus show that a single act of unprotected vaginal intercourse has a surprisingly low risk of infecting the uninfected partner--perhaps one in 100 to one in 1000. For an average, consider the risk to be one in 500. If there are 100 acts of intercourse with an infected partner, the odds of infection increase to one in five.**
>
> **Statistically, 500 acts of intercourse with one infected partner or 100 acts with five partners lead to a 100% probability of infection (statistically, not necessarily in reality).**

Assume virus transmissions in 500 acts of intercourse are mutually independent. Let:

T_i = event of transmission on the ith act

$NT_i = T_i^c$ = event of no transmission on the ith act

T = event of one or more tranmissions in 500 acts

Find P(T).

Example 1.6-3: Infectivity (continued)

P(T) =

Chapter 1: Main Points

- **Sample spaces: unions, intersections, compliments, disjoint events**
- **Probability measures: axioms, classical definition, probability of a compliment, probability of null set, addition law**
- **Counting methods: multiplication principle, permutations, combinations, binomial coefficient, multinomial coefficient**
- **Conditional probability: Multiplication law, Bayes Rule, independence of events.**

Chapter 2: Random Variables

Definition: A random variable (RV) is a function from Ω to the real numbers.

$$X: \Omega \rightarrow R^1 \quad \text{“X maps } \Omega \text{ into the reals”}$$

- **X is simply a function that assigns numbers to events. Because the events occur randomly, so does X**
- **Assigning numbers to outcomes allows us to compute minima, maxima, averages etc (i.e., to do statistics)**

2.1 Discrete Random Variables

Definition: A discrete random variable is a random variable that can take on only a finite or at most a countably infinite number of values

Definition: A continuous random variable is a random variable that can take on an uncountably infinite number of values

Example 2.1-1 A coin is thrown three times. What is Ω? Is Ω discrete or continuous?

Example 2.1-2

(See Example 2.1-1) Let X denote the number of heads that turn up for each elementary event $\omega \in \Omega$. Give $X(\omega)$ for all $\omega \in \Omega$.

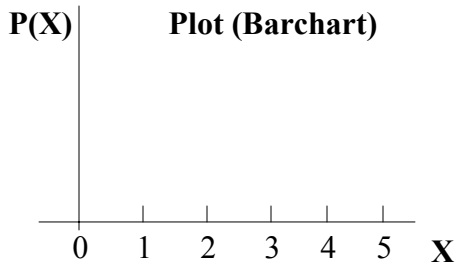
ω	$X(\omega)$

(Discrete) Probability Distribution

Definition: The probability distribution of the random variable X is a listing of all values of X and associated probabilities P(X).

Example 2.1-3 Give the probability distribution of the random variable X in Example 2.1-2.

X	P(X)



Probability and Cumulative Distribution Functions

Definition: A *probability (mass) function* of a discrete random variable X is a function p such that $p(x_i) = P(X = x_i)$ and $\Sigma_i p(x_i) = 1$.

Definition: A cumulative distribution function (CDF), F, of a random variable X is defined as:

$$F(x) = P(X \leq x)$$

Properties of a CDF:

(1) F is monotone increasing: $F(x) \leq F(x + t)$ for $t \geq 0$.

(2) Minimum is 0: $F(x) \rightarrow 0$ as $x \rightarrow -\infty$

(3) Maximum is 1: $F(x) \rightarrow 1$ as $x \rightarrow +\infty$

Example 2.1-4

Give the CDF of the random variable X in Example 2.1-3. Also plot the CDF

Independence of Random Variables

Definition: Two discrete random variables X and Y are independent if:

$$P(X = x_i \text{ and } Y = y_j) = P(X = x_i)P(Y=y_j)$$

for all i and j.

The definition can be extended to three of more random variables in the obvious way.

Well Known Discrete Distributions

We'll now turn to 5 famous discrete distributions: Bernoulli, Binomial, Geometric, Negative Binomial, and Poisson

2.1.1 Bernoulli Distribution

A coin is flipped that may or may not be balanced. $\Omega = \{H, T\}$. $P(H) = p$, $P(T) = 1 - p$.

Let X denote the number of heads observed in one flip.

X	P(X)
0	
1	

X has a Bernoulli distribution with parameter p.

Bernoulli Distribution Probability Function

Two ways to represent:

$$\text{Representation 1: } p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Representation 2: } p(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{if } x = 0 \text{ or } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

2.1.2 The Binomial Distribution

Suppose n independent Bernoulli (p) trials (or experiments) take place. The outcomes are X_1, X_2, …, X_n. The total number of "ones" or "successes" in the n trials is:

$$X = \sum_{i=1}^{n} X_i$$

Then X follows a Binomial distribution with parameters p and n.

Note: X is the number of successes in n independent Bernoulli trials. X can take on values from 0 to n.

Binomial Probability Function

Find P(X = k) where X ~ Binomial(n,p) for k = 0, 1, …, n.

(1) What is the probability of k successes followed by (n-k) failures?

(2) Is this the answer to P(X = k)?

(3) How many ways can we have k successes in n trials?

(4) So by the multiplication principle, P(X = k) is:

Example 2.1-5 Freethrow Shooting

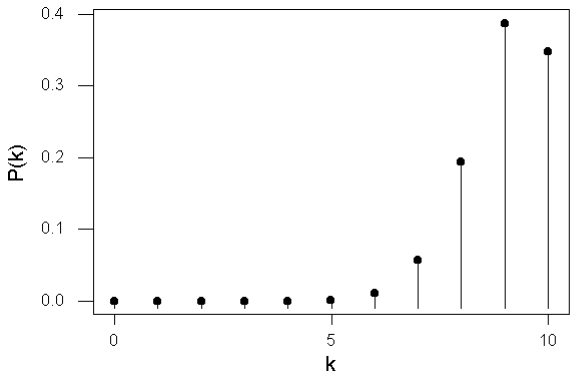
Suppose you are a great freethrow shooter in basketball. You hit 90% of your shots on average. Find the probability that you make:

(a) All 10 of your next 10 shots

(b) 9 out of your next 10 shots

(c) 5 out of your next 10 shots

Binomial (10, .9) Probability Mass Function

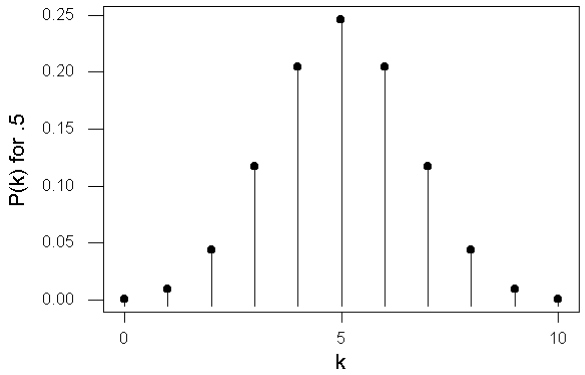


Binomial "Symmetry"

Show that if p = .5, the Binomial probability function is symmetric:

$$p(k) = p(n-k)$$

Binomial (10, .5) Probability Mass Function



Geometric Distribution: Geom(p)

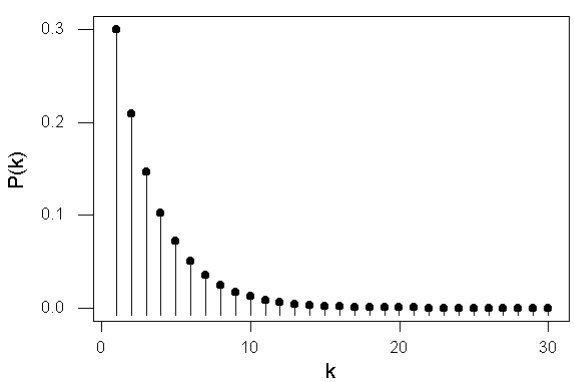
Also for a sequence of Bernoulli (p) trials, let X denote the total number of trials that it takes to get one success. For X = k, we must have k-1 failures followed by 1 success:

$$P(X = k) = P(X_1 = 0) \times P(X_2 = 0) \times \cdots \times P(X_{k-1} = 0) \times P(X_k = 1)$$
$$=$$

Fun Fact: $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$, for $0 < x < 1$.

Example 2.1-6 Using the above fun fact, show that $\sum_{k=0}^{\infty} p(k) = 1$.

Geometric (.3)



Negative Binomial and Hypergeometric Distributions

Negative Binomial: NB(p,r). **Just like the geometric distribution, except X is the number of trials until there are r failures.**

P(X = k) = P(r-1 successes in k-1 trials)P(Success on kth trial)

=

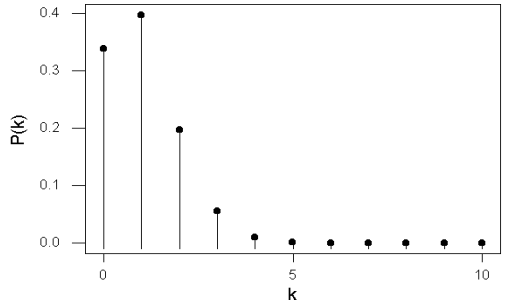
2.1.4 Hypergeometric Distribution: H(n,r,m).

Have n equally likely items in Ω, with r items of type 1, n - r items of type 2. Sample m items without replacement. X is the number of items of type 1. (We've seen this already.)

P(X = k) = p(k) =

Example 2.1-7 (Quality control acceptance sampling)

From Example 1.4-10: You receive product in lots of size 200. Suppose 20 of these items are defective, and 180 are not defective. Choose 10 items at random for inspection. Distribution looks as follows:



2.1.5 Poisson Distribution

Used to model probabilities of counts (0, 1, 2, …) over fixed units of time or space:

- **Number of customers arriving at a store for a unit time**
- **Hits on a website over a unit of time**
- **Number of cancers annually of a specific type in a population**
- **Number of lightening strikes in fixed units of land (e.g., square miles)**
- **Number of bids received in an auction for offshore oil tracts of land**

Poisson Probability Function

If a random variable X follows a Poisson distribution with parameter λ (X ~ Pois(λ)), then P(X = k), for k = 0, 1, 2, … is given by the Poisson probability function:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \ k = 0, 1, 2, \ldots$$

Fun Fact: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$.

Example 2.1-8 Using the above fun fact, show that $\sum_{k=0}^{\infty} p(k) = 1$.

Poisson distribution can be developed as a limiting form of the binomial distribution

Consider the number of auto deaths per year in a rural county. Break the year into small time intervals, say minutes: 365 days x 24 hrs/day x 60 min/hr = 525,600 minutes

On average we get:

$$\lambda = \frac{4 \text{ deaths}}{\text{year}} = \frac{4 \text{ deaths}}{525{,}600 \text{ trials}}$$

So p = probability of death on any given trial (minute) is 4/525,600 = .00000761.

Poisson Development (continued)

Then p(k) should be approximately binomial with n = 525,600 and p = .00000761, as long as the binomial assumptions hold approximately:

(1) Intervals are small enough so that the probability of TWO OR MORE events in an interval is negligible

(2) The probability of an event in any two subintervals is the same (probability of a success is constant across the year)

(3) Whatever occurs in one subinterval is independent of what occurs in the other

Poisson as Limit of Binomial

$$\text{Fun Fact: As } n \to \infty,\ (1 - \frac{\lambda}{n})^n \to e^{-\lambda}$$

Use the above fun fact to show that as:

$$n \to \infty \text{ and } p \to 0 \text{ such that } np = \lambda,$$

the binomial probability function converges to the Poisson probability function.

Poisson as Limit of Binomial (continued)

Example 2.1-9

Use Minitab to compare the Poisson probabilities for $\lambda = 4$ to the binomial probabilities for n = 365 and $p = \lambda / n$ = .0109589, for k = 0, 1, 2, …, 10.

k	P(k) Bin	P(k) Pois
0	0.018	0.018
1	0.072	0.073
2	0.146	0.147
3	0.196	0.195
4	0.196	0.195
5	0.157	0.156
6	0.104	0.104
7	0.059	0.060
8	0.029	0.030
9	0.013	0.013
10	0.005	0.005

Check, for k = 1 using Minitab:

```
MTB > let k2 = 365*k1*(1-k1)
MTB > print k2
```

Data Display

```
K2    3.95616
MTB > let k2 = 365*k1*(1-k1)**364
MTB > print k2
```

Data Display

```
K2    0.0724568
MTB > let k2 = 4*exp(-4)
MTB > print k2
```

Data Display

```
K2    0.0732626
```


Example 2.1-10 Fatalities from Horse Kicks

Bortkiewicz (1898) documented fatalities from horse kicks in the Prussian army over a 20-year period. There were 10 corps followed giving 200 corps-years of data. The total number of deaths was 122, so $\lambda = 122/200 = .61$. How well would a Poisson(.61) distribution fit?

Number of deaths per year, k	Observed Frequency (of each k)	Observed Relative Frequency	Poisson Probability
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003
Totals:	200	1.000	0.999

2.2 Continuous Random Variables

Suppose X is a random variable that takes on a continuum of values.

Example 2.2-1 Suppose we install a new light bulb and let X denote the time to failure.

(a) Find Ω.

(b) Find the probability that the bulb fails at precisely 78.000000…. hours.

Lesson: For a continuous random variable P(X = x) =

Probability Density Functions (PDFs)

We instead consider the probability that the bulb fails in an interval [a,b] with b > a. This probability can be positive. To obtain these "interval" probabilities, we use the probability density function, f

Definition: A piecewise continuous function f, such that

$$(1) \quad f(x) \geq 0, \qquad -\infty < x < \infty$$
$$(2) \quad \int_{-\infty}^{\infty} f(x)dx = 1$$

is a *probability density function*. If a random variable X has density f, then

$$P(a < X < b) = \int_a^b f(x)dx$$
$$= \text{"area under the curve from } a \text{ to } b\text{"}$$

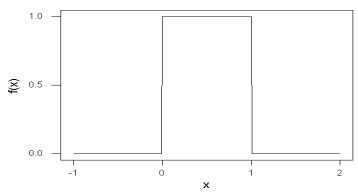
Uniform Distribution on [0,1]

If we say "choose a number at random between 0 and 1," we mean: choose a number with all intervals of the same length being equally likely. This leads to the density:

Uniform density on [0,1]:

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Picture:



Example 2.2-2

X ~ Uniform on the interval [0,1] (i.e., X ~ Unif(0,1)). Using geometry, find:

a) $P(X \le .5)$

b) $P(.75 < X \le 1)$

c) $P(0 < X < 1)$

d) $P(X = .5)$

Example 2.2-2 (continued)

X ~ Uniform on the interval [0,1] (i.e., Unif(0,1)). Using calculus, find:

a) $P(X \leq .5)$

b) $P(.75 < X \leq 1)$

c) $P(0 < X < 1)$

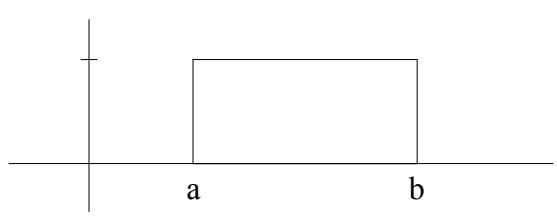
d) $P(X = .5)$

Uniform Density on [a,b]

If X is distributed as a uniformly on the interval [a,b], then the density f is given by:

$$f(x) = \begin{cases} \quad , & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

Picture:



Cumulative Probability Distribution, F

The *cumulative distribution function* (CDF) for a continuous random variable X is given by:

$$F(x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(u)du$$

Note 1: If f is continuous at x, then:

$$f(x) = \frac{d}{dx}F(x) = F'(x)$$

Note 2: Probability of an interval can be obtained from F:

$$\begin{aligned} P(a < X < b) &= \int_a^b f(x)dx \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\ &= F(b) - F(a) \end{aligned}$$

Example 2.2-3

Suppose:

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

a) Plot f and determine if f is a density function.

b) If f is a density, find the CDF, F.

C) Find P(.25 < X < .5)

Quantiles, Percentiles, and the Inverse CDF, F^{-1}

Definition: X is a random variable with CDF F. Then the pth *quantile* is the value x such that F(x) = p, for 0<p<1.

<u>Note 1</u>: The pth quantile is sometimes called the 100(p)th percentile.

<u>Note 2</u>: Suppose F is strictly increasing ($b > a \Rightarrow F(b) > F(a)$) and p = F(x) (x is the pth quantile). Then often x can be found by solving for x in terms of p (by inverting F):

$$x = F^{-1}(p)$$

<u>Note 3</u>: Certain quantiles (percentiles) have special names:

Quantile	Percentile	Special Name
0.25	25th	First quartile
0.75	75th	Third quartile
0.5	50th	Median (Second quartile_)
0.9	90th	9th decile
0.1	10th	First decile

Example 2.2-4

a) For the CDF in Example 2.2-3(b), find $F^{-1}(p)$

b) Using (a), find the median , the lower (first) quartile and the 99th percentile.

Example 2.2-5

Suppose $F(x) = x^3$, for $0 \le x \le 1$, and 0 otherwise.

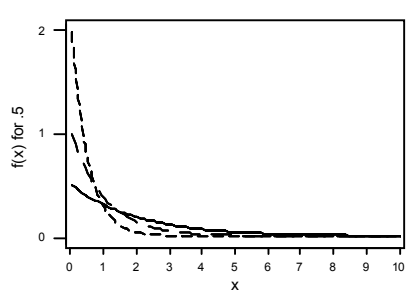
a) Find F^{-1}

b) Find the density function, f.

2.2.1 The Exponential Density: X~Exp(λ)

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

Plots of *f* for λ = .5, 1, and 2:



Example 2.2-6

a) Find the CDF of the exponential distribution with parameter λ.

b) Find F^{-1}

c) Find the median.

Memoryless Property of Exponential

A distribution is memoryless if:

$$P(X > x) = P(X > x + s \mid X > s)$$

Example: Let X denote the time to failure of an electronic component. Many such components have life spans are "memoryless." If the component has lasted s = 2 years, the probability that it lives 2 more years is the same as if it had just been installed!

Example: Consider human mortality: X is the age in years at which a human dies. Let x = 15 and s = 70. Do you think that:

$$P(X > 15) = P(X > 15 + 70 \mid X > 70)?$$

Example 2.2-7

Show that the an exponential random variable is memoryless.

Notes on Exponential:

(1) The exponential distribution is the only memoryless distribution!

(2) If X ~ Pois (λ), the distribution of time between arrivals is Exp(λ)

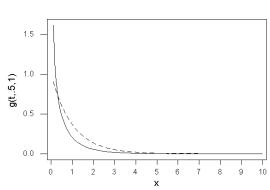
2.2.2 The Gamma Density

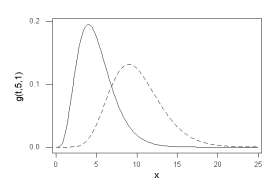
A random variable T follows a gamma distribution with shape parameter α and scale (width) parameter λ if the density is given by:

$$g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} \qquad t > 0$$

where:

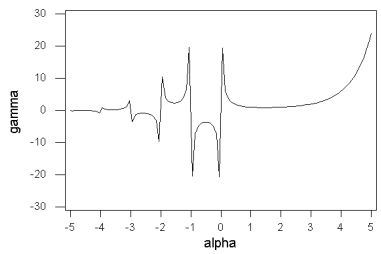
$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$





The Gamma Function

1) No general analytic solution to this integral
2) $\Gamma(1) = 1$
3) $\Gamma(.5) = \sqrt{\pi}$
4) $\Gamma(x+1) = x\,\Gamma(x)$
5) $\Gamma(n) = (n - 1)!$ $n = 1, 2, 3, \ldots$



Why Bother with Gamma Distribution?

1) Sometimes used as a model for data. (Although this is pretty rare.) Text author used it to model interarrival times for a sequence of small earthquakes in California. (I guess the exponential wasn't quite flexible enough to fit the data.)

2) Two famous, useful distributions, the exponential (which we have seen) and the chi-square distribution (which we will see later) are special cases:

   - Gamma(1, λ) = Exp(λ)
   - Gamma(n/2, .5) = chi-square(n)

The Normal (or Gaussian) Distribution

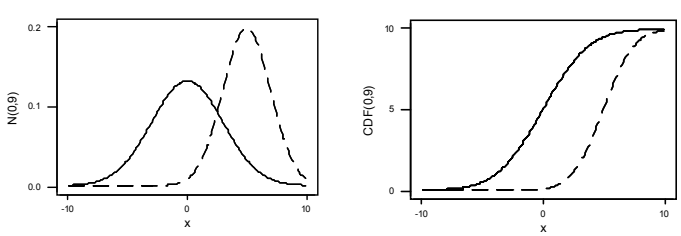
- **Proposed by Carl Friedrich Gauss**
- **Used to model:**
  - **measurement errors**
  - **persons' heights**
  - **many physical phenomena**
  - **IQ scores**
  - **Grading/job performance (inappropriately)**
  - ***Any quantity that is the sum of a large number of independent random variables***

Normal Density and CDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}[\frac{x-\mu}{\sigma}]^2} \quad -\infty < x < \infty$$

Normal CDF can't be integrated analytically; use numerical methods to approximate.

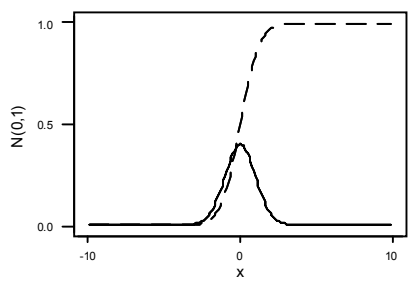
Densities and CDFs for N(0,9) and N(5,4):



The Standard Normal Distribution

Very special case: When $\mu = 0$ and $\sigma = 1$, we obtain the standard normal density $\phi(x)$ and CDF $\Phi(x)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad -\infty < x < \infty$$



2.3 Functions of a Random Variable

Suppose we know the density and CDF of X, which we will call f_X and F_X. But we're really interested in a function of X, Y = g(X). What are the density and CDF of Y?

- **Suppose X is return on a stock. What is the distribution of Y = g(X) = ln(X) (log returns)?**
- **Suppose X = mileage for a new car in miles per gallon, and suppose X ~ N(20,1). What is the distribution of Y = 1/X = gallons per mile?**
- **Suppose X = velocity of a particle of mass m. What is the pdf of the particle's kinetic energy $Y = .5mX^2$?**
- **Suppose f_X is the pdf of X, where X is the daytime high in ^oF for September 25 in Minneapolis. What is the PDF of Y = High in ^oC = (5/9)(X - 32).**

Functions of a Random Variable

We will develop formulas for the PDFs for three functions of X:

(1) Y = aX + b, where a and b are constants (linear)

(2) $Y = X^2$ (square)

(3) Y = 1/X (inverse)

The method for doing this is as follows:

Step 1: Find CDF $F_Y(y) = P(Y < y)$ using the CDF of X

Step 2: Differentiate F_Y to get the PDF of Y, f_Y.

Function 1: Let Y = aX + b

Assume X has differentiable CDF F_X and PDF f_X. Then (for appropriate values of y, and $a \neq 0$):

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

$$F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right) & a < 0 \end{cases}$$

Justification: (a > 0)

Step 1:

Step 2:

Function 1: Let Y = aX + b (continued)

Justification: (a < 0)

Step 1:

Step 2:

Example 2.3-1

Suppose X ~ N(μ,σ²). Find the PDF of :

$$Z = \frac{X - \mu}{\sigma} = aX + b$$

where $a = \frac{1}{\sigma}$ and $b = \frac{-\mu}{\sigma}$.

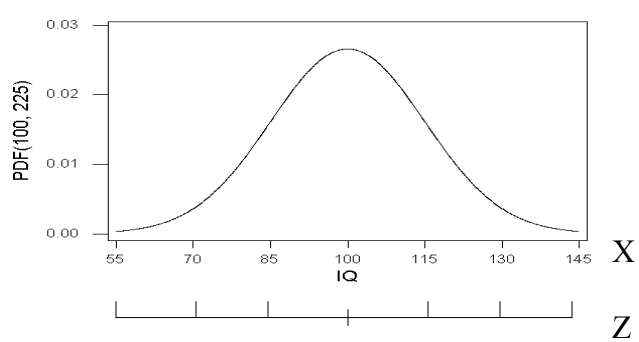
Consequence of Example 2.3-1

$$F_X(x) = P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma})$$
$$= P(Z \le \frac{x-\mu}{\sigma})$$
$$= \Phi(\frac{x-\mu}{\sigma})$$

- **Implication: One can find cumulative probabilities for any N(μ,σ^2) random variable using the standard normal cdf . (Only need one table).**

Example 2.3-2

(a) Suppose IQ (intelligence quotient) scores follow a distribution that is N(100,15^2). Find P(X < 85).



Example 2.3-2 (continued)

(b) Find P(75 ≤ X ≤ 125)

(c) What percentage of the population is within 1, 2, and 3 standard deviations of the mean?

Function 2: Let $Y = X^2$

Assume X has differentiable CDF F_X and PDF f_X. Then (for appropriate values of y ≥ 0):

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$f_Y(y) = \frac{f_x(\sqrt{y})}{2\sqrt{y}} + \frac{f_x(-\sqrt{y})}{2\sqrt{y}}$$

Step 1:

Step 2:

Example 2.3-3

Let $Y = Z^2$, where Z is a standard normal random variable. Find $f_Y(y)$.

Note 1: $Y = Z^2$ is a chi-square RV with 1 degree of freedom

Note 2: This is also a gamma density with $\alpha = \lambda = .5$.

Function 3: Let Y = 1/X

Assume X has differentiable CDF F_X and PDF f_X. Then (for appropriate values of y):

$$F_Y(y) = 1 - F_X(\frac{1}{y})$$

$$f_Y(y) = \frac{f_X(y^{-1})}{y^2}$$

Step 1:

Step 2:

Proposition B

Let X be a continuous random variable with density f(x) and Y = g(X) where g is a differentiable, strictly montonic function on some interval I. Suppose that f(x) = 0 if x is not in I. Then Y has the density function:

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$$

for y such that y = g(x) for some x, and $f_Y(y) = 0$ if $y \neq g(x)$ for any x in I.

Example 2.3-4: Let Y = 1/X (where X is always positive: $f_X(x) = 0$, for $x \leq 0$)

Applying Proposition B directly:

(1) Y = g(X) = 1/X (strictly decreasing for X > 0)

(2) Solve for X to get g^{-1}:

$X = 1/Y = g^{-1}(Y)$

(3) Compute derivative of g^{-1}: $\frac{d}{dy}g^{-1}(y) = \frac{-1}{y^2}$

(4) Plug into formula:

$$f_Y(y) = f_X(\frac{1}{y})|\frac{-1}{y^2}|$$
$$= \frac{f_X(y^{-1})}{y^2}$$

(compare with p.110)

How to Simulate a Distribution: Generating Random Numbers from a CDF F

Suppose U is uniform on [0,1] and let $R = F^{-1}(U)$. How is R distributed?

$$P(R < r) = P(F^{-1}(U) < r) = P(U < F(r)) = F(r)$$

So, to generate random numbers R that follow a distribution with invertible CDF F:

(1) Generate U from Uniform[0,1]

(2) Compute $R = F^{-1}(U)$.

Chapter 3: Joint Distributions

- **Discrete example: Suppose**

  X = number of wolves in St. Louis County, Minnesota; probability function is $p_X(x)$.

  Y = number of deer in St. Louis County; probability function is $p_Y(y)$.

  What is the joint distribution, $p_{XY}(x,y)$?

- **Continuous example: Suppose**

  X = person's height, with density $f_X(x)$.

  Y = person's weight, with density $f_Y(y)$.

  What are the joint PDF and CDF, $f_{XY}(x,y)$ and $F_{XY}(x,y)$?

3.2 Discrete Random Variables

Example 3.2-1 A Coin is tossed 3 times :

Ω = {ttt,htt,tht,tth,hht,hth,thh,hhh}

X = Number of heads on first toss

Y = Total number of heads

	Y				
X	**0**	**1**	**2**	**3**	**Totals:**
0					
1					
Totals:					

***Joint Probability Function*: $P_{XY}(x_i, y_j) = P(X = x_i, Y = y_j)$**

- **$P_{XY}(1, 0) =$**
- **$P_{XY}(0, 2) =$**

Marginal Probability Functions

Probability of an event A:

$$P((X,Y) \in A) = P(A) = \sum_{(x_i,y_i)\in A} P_{XY}(x_i, y_i)$$

Marginal Probability Functions:

$$P_X(x_i) = \sum_j P_{XY}(x_i, y_j)$$

$$P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

For 3 variables, etc.: $P_{XY}(x_i, y_j) = \sum_k P_{XYZ}(x_i, y_j, z_k)$

3.3 Continuous Random Variables

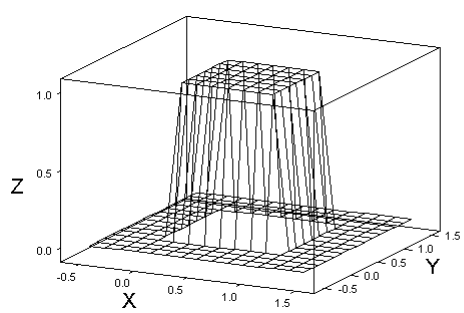
- **With one continuous RV, probability was the area under the density curve between a and b.**
- **With two jointly distributed RVs X and Y, probability is the volume under the density curve over the region A in XY space (total volume = total probability = 1):**

$$P((X,Y) \in A) = P(A) = \int\int_{(X,Y)\in A} f_{XY}(x,y)dxdy$$

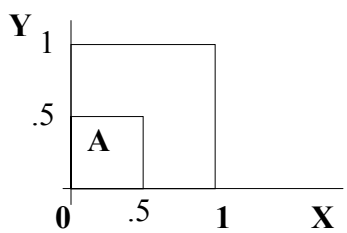
- **Example 3.2-2 Suppose X and Y have a joint uniform distribution on [0,1]x[0,1]. Find P(X ≤ .5, Y≤ .5):**

$$P((X,Y) \in A) = P(A) = \int_0^{.5}\int_0^{.5} f_{XY}(x,y)dxdy$$

Picture: Bivariate Uniform Distribution on [0,1]x[0,1]:



Region A:

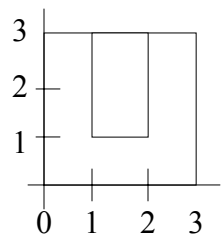


Example 3.3-3

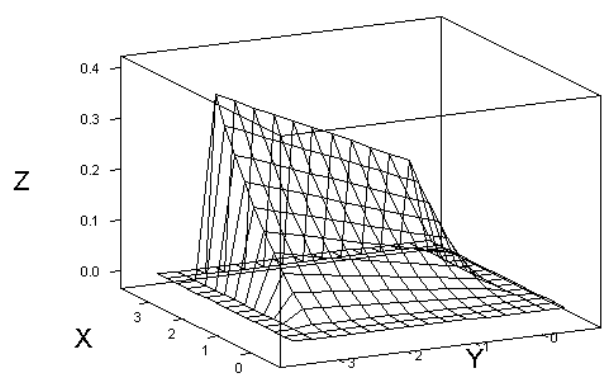
Suppose:

$$f_{XY}(x,y) = \begin{cases} \frac{4}{189}(x^2 + xy), & 0 \le x \le 3, \qquad 0 \le y \le 3 \\ 0, & \text{otherwise} \end{cases}$$

Find P(A), where A = {(x,y)|1≤x ≤ 2,1 ≤ y ≤ 3}



Picture



Using some integration:

$$P((X,Y) \in A) = \int_1^3 \int_1^2 \frac{4}{189}(x^2 + xy)dxdy$$

CDF vs PDF

Cumulative Distribution Function:

$$\begin{aligned} F_{XY}(x,y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x,y)dxdy \end{aligned}$$

Probability Density Function:

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$$

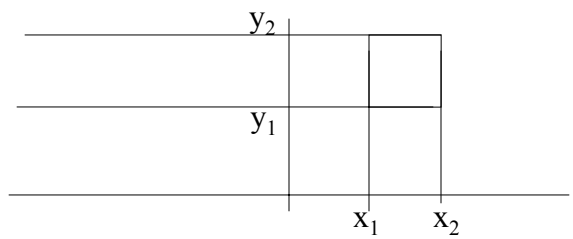
Note on CDF

Recall that for the univariate case that:

$P(a \le X \le b) = F(b) - F(a)$

In the multivariate case, not quite so simple:

$$\int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y)dxdy \neq F(x_2,y_2) - F(x_1,y_1)$$



$$P((X,Y) \in A) = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1)$$

Marginal Density Function

- **Marginal density:**

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

- **Marginal CDF**

$$F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(u,y)dydu$$

Example 3.3-4

(a) Find the marginal density of X for the density in Example 3.3-3.

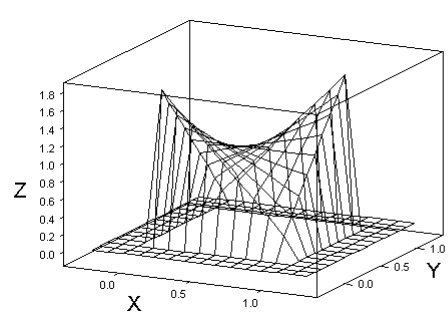
Example 3.3-4 (continued)

(b) Find the marginal density of Y

Quiz: Suppose we know that the marginal densities of X and Y are Uniform[0,1]. Is the joint density uniform?

Counterexample

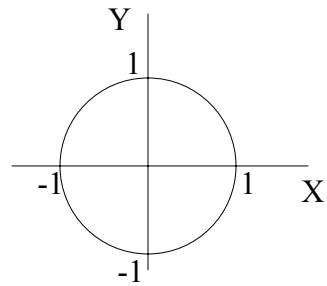
Consider the density: f(x,y) = 2 - 2x - 2y + 4xy on [0,1]x[0,1]. Both X and Y have uniform marginal distributions



Example 3.3-5

Assume that X and Y are uniformly distributed over the unit circle:

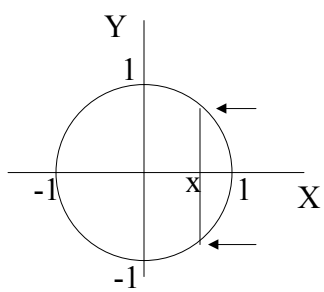
$$f_{XY}(x,y) = \begin{cases} \quad , \\ 0, \quad \text{elsewhere} \end{cases}$$



Find the marginal density of X. (Only trick here is to have the right limits of integration over y!)

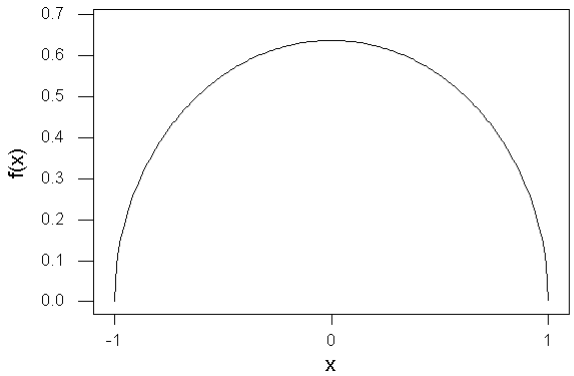
Example 3.3-5 (continued)

Picture:



$f_X(x)$ =

Picture of Marginal Density



Why is it "fat" in the middle, when the joint density was uniform?

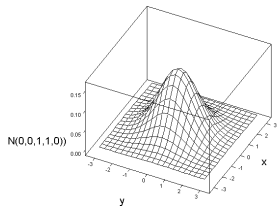
The Bivariate Normal Distribution

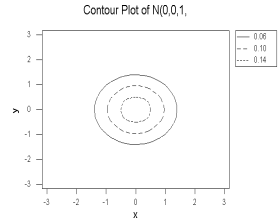
(X,Y) ~ N(μ_x,μ_y,σ_x^2,σ_y^2,ρ) if:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

$$exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)$$

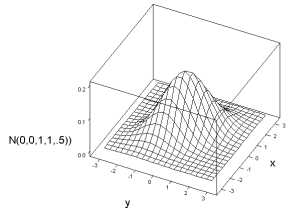
for $-\infty < x, y < \infty$, $-1 < r < 1$, μ_x, μ_y, σ_x, σ_y positive.

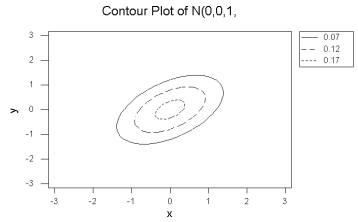




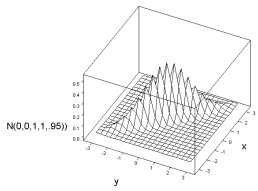
More Normal Distributions

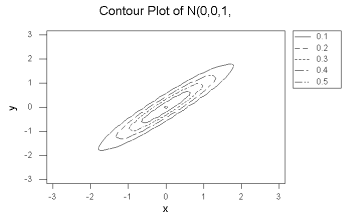
N(0,0,1,1,.5):





N(0,0,1,1,.95):





BV Normal Marginals

Fact: If (X,Y) ~ N(μ_x,μ_y,σ_x^2,σ_y^2,ρ), then:

X~N(μ_x,σ_x^2) and Y ~ N(μ_y,σ_y^2).

Proof: By "simply" carrying out the integration:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

This requires a substitution trick followed by a "completing the square" trick (see text). In the end, it leads to a univariate normal density as described.

3.4 Independent Random Variables

Definition: The random variables $X_1, X_2, \ldots, X_n$ (discrete or continuous) are said to be *independent* if their joint CDF factors into the product of their marginal CDFs:

$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$$

for all $x_1, x_2 \ldots, x_n$.

Equivalently, random variables are independent if and only if their joint density (probability) function factors:

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(u) f_Y(v) dv du \\ &= \left[\int_{-\infty}^{x} f_X(u) du \right] \left[\int_{-\infty}^{y} f_Y(v) dv \right] \\ &= F_X(x) F_Y(y) \end{aligned}$$

Example 3.4-1

Show that if $(X,Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ and $\rho = 0$, then X and Y are independent random variables.

3.5 Conditional Distributions

Example 3.5-1 Recall Example 3.1-1. A Coin is tossed 3 times: X = #heads on first toss, Y = Total number of heads

	Y				
X	**0**	**1**	**2**	**3**	**Totals:**
0	1/8	2/8	1/8	0	0.5
1	0	1/8	2/8	1/8	0.5
Totals:	1/8	3/8	3/8	1/8	1

The conditional probabilities of X given Y = 1 are:

- $P_{X|Y}(0|1) =$
- $P_{X|Y}(1|1) =$

3.5.1 Discrete Case

The conditional probability that $X = x_i$ given $Y = y_j$, if $p_Y(y_j) > 0$:

$$\begin{aligned} p_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\ &= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \\ &= \frac{p_{XY}(x_i, y_j)}{p_Y(y_j)} \end{aligned}$$

$p_{Y|X}(x_i|y_j)$ is defined to be zero if $p_Y(y_j) = 0$.

Law of Total Probability:

$$p_X(x) = \sum_y p_{X|Y}(x|y) p_Y(y)$$

3.5.2 Continuous Case

The conditional density of X given Y, if $f_Y(y) > 0$, is:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Note that this is consistent with differential argument:

$$P(x \le X \le x + dx | y \le Y \le y + dy) = \frac{f_{XY}(x,y)dxdy}{f_Y(y)dy}$$
$$= \frac{f_{XY}(x,y)}{f_Y(y)}dx$$

Law of Total Probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$$

Example 3.5-2

Suppose (X,Y) ~ N(μ_x,μ_y,σ_x^2,σ_y^2,ρ). Then the conditional distribution of Y|X is univariate normal with mean and variance:

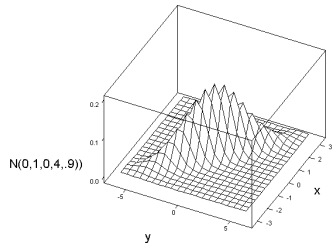
$$\mu_{Y|X}(y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$
$$\sigma^2_{Y|X} = \sigma^2_Y(1 - \rho^2)$$

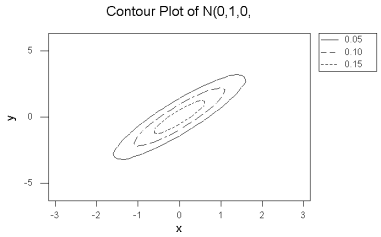
Notes:

(1) $\mu_{Y|X}(y|x)$ is the regression function of Y given X. Slope is $\rho(\sigma_Y/\sigma_X)$.

(2) The variance of Y given X = x is constant (does not depend on x).

Picture, for $\sigma_X = 1$, $\sigma_Y = 2$, $\rho = .90$:

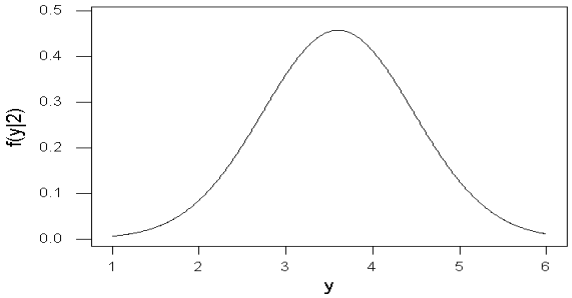




Conditional Distribution at x = 2

Mean: 0 + .9(2/1)(2-0) = 3.6

Variance: $2^2(1-.9^2) = 4(.19) = .76$, so $\sigma = .87$



3.6 Functions of Jointly Distributed Random Variables

Objective: Find PDFs, CDFs of a funtion of two or more random variables.

We will consider two important special cases and then give a general method:

(1) Sums: Z = g(X,Y) = X+Y

(2) Quotients: Z = g(X,Y) = X/Y

(3) General method

3.6.1 Sums and Quotients

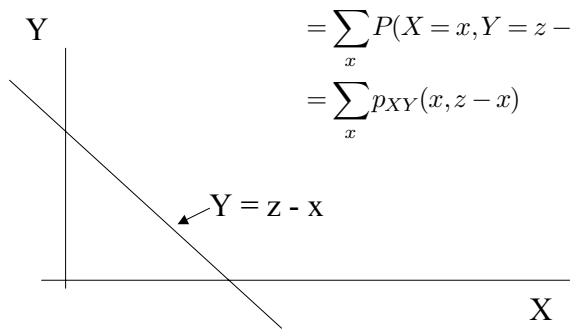
Let Z = X + Y

Discrete Case:

$$P(Z=z) = \sum_{(x,y)|x+y=z} P(X=x, Y=y)$$

$$= \sum_x P(X=x, Y=z-x)$$

$$= \sum_x p_{XY}(x, z-x)$$



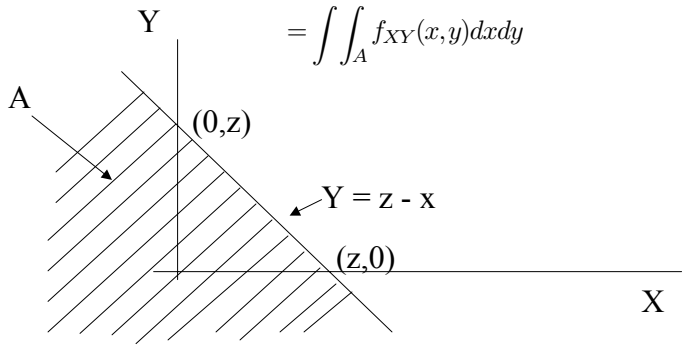
Continuous Case for Z = X + Y

Recall steps:

(1) Find $F_Z(z)$ in terms of $F_{XY}(x,y)$

(2) Differentiate to find $f_Z(z)$.

$$F_Z(z) = P(Z \le z) = P(X + Y \le z)$$
$$= \int\int_A f_{XY}(x,y)dxdy$$



Carrying out integration over A:

First make change of variable: y = v - x, so:

- **dy/dv =1 yields dy=dv**
- **Limits: When y = -∞, v = -∞; when y = z-x, v = z**

$$F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_{XY}(x, v-x)dvdx$$

Change order of integration and differentiate to get f:

$$F_Z(z) = \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_{XY}(x, v-x)dxdv$$
$$F'(z) = f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z-x)dx$$
$$= \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx \qquad \text{if X, Y independent}$$

(Compare to discrete result--obvious analogy)

Justification of Result in Example 3.5-2

Example 3.6-1

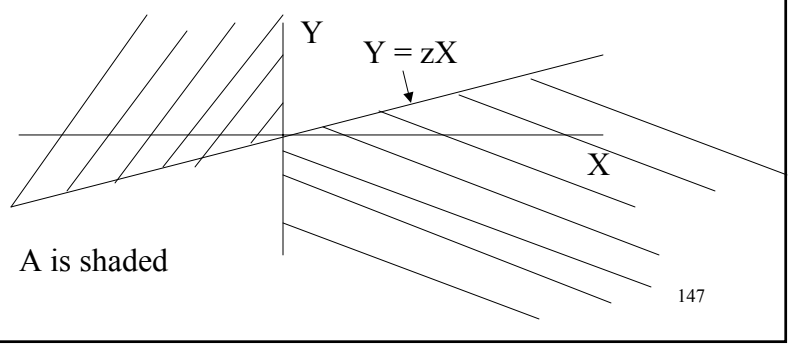
The lifetime T_1 of an electronic component is Exp(λ). This component has an identical backup component that goes into service if the first component fails. The lifetime T_2 of the backup this is also Exp(λ). Find the distribution of the lifetime of the system: $Z = T_1 + T_2$.

Quotients: Z = Y/X

As always:

$$F_Z(z) = P(\{(x,y)|y/x \leq z\}) = P(A)$$

- **When x is positive, $y \leq zx$**
- **When x is negative, $y \geq zx$**



Carrying out the integration:

$$F_Z(z) = \int_{-\infty}^{0} \int_{xz}^{\infty} f_{XY}(x,y)dydx + \int_{0}^{\infty} \int_{-\infty}^{xz} f_{XY}(x,y)dydx$$

First make change of variable: y = vx, so:

- **dy/dv =x yields dy=xdv**
- **Limits:**

  (x < 0): When y = zx, v = z; when y = ∞, v = -∞;

  (x > 0): When y = -∞, v = -∞; when y = zx, v = z

Integration (continued)

$$
\begin{aligned}
F_Z(z) &= \int_{-\infty}^{0}\int_{z}^{-\infty} x f_{XY}(x, xv)dvdx + \int_{0}^{\infty}\int_{-\infty}^{z} x f_{XY}(x, xv)dvdx \\
&= \int_{-\infty}^{0}\int_{-\infty}^{z} (-x) f_{XY}(x, xv)dvdx + \int_{0}^{\infty}\int_{-\infty}^{z} x f_{XY}(x, xv)dvdx \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} |x| f_{XY}(x, xv)dxdv \\
&= \int_{-\infty}^{\infty} |x| f_{XY}(x, xz)dx
\end{aligned}
$$

If X and Y are independent,

$$f_Z(z) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz)dx$$

Example 3.6-2: Cauchy Distribution

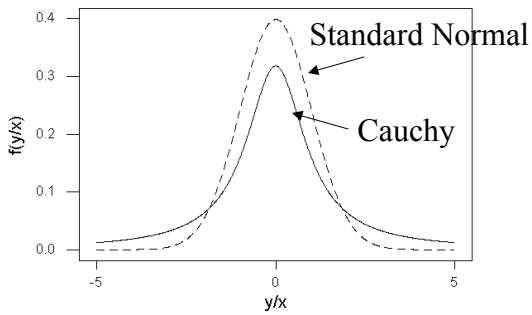
Suppose X and Y are independent standard normal random variables. Let Z = Y/X and find the density of Z.

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} \frac{|x|}{2\pi} e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}(xz)^2} dx \\
&= 2\int_{0}^{\infty} \frac{x}{2\pi} e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}(xz)^2} dx
\end{aligned}
$$

Substitution: u = x^2; then du/dx = 2x; du = 2xdx:

Cauchy (continued)

- **Note on Cauchy: Shaped like normal, but has heavy tails--Z blows up if X in Z = Y/X is near zero!**
- **This leads (we shall see) to an infinite variance**



General Method (Proposition A in text)

New random variables U and V are differentiable functions of the old random variables X and Y:

$$u = g_1(x,y) \qquad v = g_2(x,y)$$

Assume it is possible to solve for x and y in terms of u and v:

$$x = h_1(u,v) \qquad y = h_2(u,v)$$

Assume the Jacobian is nonzero for all x and y:

$$J(x,y) = \det \begin{bmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{bmatrix} = \left(\frac{\partial g_1}{\partial x}\right)\left(\frac{\partial g_2}{\partial y}\right) - \left(\frac{\partial g_2}{\partial x}\right)\left(\frac{\partial g_1}{\partial y}\right)$$

Then:

$$f_{UV}(u,v) = f_{XY}(h_1(u,v), h_2(u,v))|J^{-1}(h_1(u,v), h_2(u,v))|$$

Example 3.6-3

Suppose that X_1 and X_2 are independent standard normal random variables, and $Y_1 = X_1$ and $Y_2 = X_1 + X_2$. Find the joint density of Y_1 and Y_2.

Extrema and Order Statistics

Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with common CDF F and density f. Let:

$X_{(1)}$ = minimum of $\{X_1, X_2, \ldots, X_n\}$
$X_{(2)}$ = 2nd smallest of $\{X_1, X_2, \ldots, X_n\}$
$X_{(3)}$ = 3rd smallest of $\{X_1, X_2, \ldots, X_n\}$
…
$X_{(n)}$ = maximum $\{X_1, X_2, \ldots, X_n\}$

Find: (a) Distribution of the minimum, $X_{(1)}$
(b) Distribution of the maximum. $X_{(n)}$
(c) Distribution of the kth smallest, $X_{(k)}$

(a) Minimum

$$
\begin{aligned}
F_{X_{(1)}} &= P(X_{(1)} \le x) = P(\text{At least one } X_i \le x) \\
&= 1 - P(\text{No } X_i \le x) \\
&= 1 - P(X_1 \ge x \cap X_2 \ge x \cap \cdots \cap X_n \ge x) \\
&= 1 - [1 - F(x)]^n
\end{aligned}
$$

Differentiating to get f:

(b) Maximum

(c) Distribution of kth order statistic

Heuristic derivation:

$$
\begin{aligned}
f_{X_{(k)}}(x) &= P(x \le X_{(k)} \le x + dx) \\
&= P(\text{First } k-1 \text{ are} < x \text{ and last } n-k \text{ are} > x + dx) \\
&= \binom{n}{k-1}\binom{n-(k-1)}{1}\binom{n-k}{n-k}[F(x)]^{k-1} f(x)[1-F(x)]^{n-k} \\
&= \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1} f(x)[1-F(x)]^{n-k}
\end{aligned}
$$

Example 3.6-4

Suppose $X_i \sim$ Unif[0,1], for i = 1, 2, …, 10. Find and plot

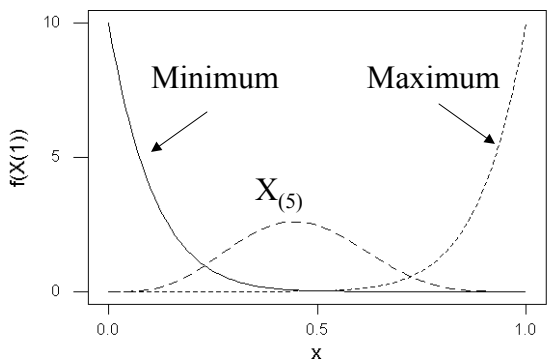
a) Density of $X_{(1)}$

b) Density of $X_{(5)}$

c) Density of $X_{(10)}$

Recall: F(x) = x, for $0 \le x \le 1$ (F(x) = 0, x < 0; F(x) = 1, x>0)

Example 3.6-3 (Continued)



Chapter 4: Expected Values

4.1 Expected value of a random variable

(1) Long run average value of X if you repeat the experiment infinitely may times.

(2) Weighted average of all possible values of X--where the weights are the corresponding probabilities of X.

(3) Center of mass of the frequency function.

Definition: If X is discrete with probability function p(x), the expected value of X is:

$$E(X) = \sum_i x_i p(x_i)$$

provided $\Sigma|x_i|p(x_i) < \infty$. E(X) undefined if the sum diverges!

Examples

Example 4.1-1 Suppose you roll a single die. Let X denote the score. Find E(X).

Example 4.1-2 Should you play the following game? Cost is \$1. You roll two dice. If the sum is 2, you win \$10. If the sum is greater than 10, you win \$5. Hint: Let Z be the amount you win or lose, and find E(Z).

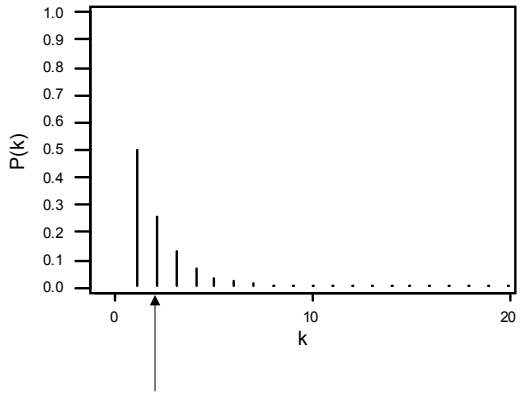
Examples

Example 4.1-3 Find E(X) if X ~ Bernoulli(p).

Example 4.1-4 Find the expected value of K if

K ~ Geom(p)

Picture of Geometric Distribution (p = .5)



Balance point is 1/p = 2

Examples

Example 4.1-5 Find the expected value X if X ~ Pois(λ).

Expectation for a Continuous RV

Definition: If X is continuous with probability density function f(x), the expected value of X is:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

provided $\int |x| f(x) dx < \infty$. E(X) undefined if $\int |x| f(x) dx \to \infty$.

Example 4.1-6 Let X ~ Unif(0,1). Find E(X).

Example 4.1-7

Suppose:

$$f(x) = \begin{cases} \frac{3}{2}x^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Find E(X)

Example 4.1-8: Normal Density

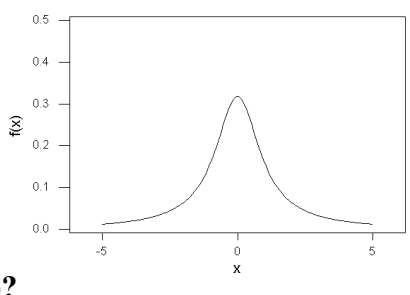
Suppose X ~ N(μ,σ^2). Find E(X).

Example 4.1-9: Cauchy Density

Recall that the density of a ratio x of standard normals is:

$$f(x) = \frac{1}{\pi}\left(\frac{1}{1+x^2}\right) \qquad -\infty \ \leq \ x \ \leq \ \infty$$

Picture:



What is E(X)?

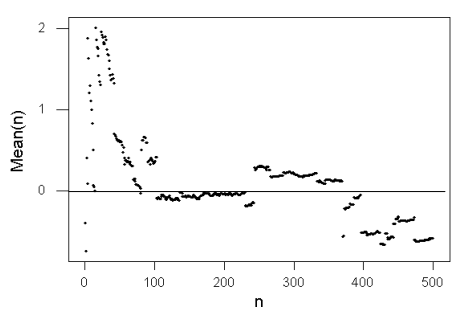
Example 4.1-9: Cauchy Density

Does E(X) exist? No!!! Let u = x^2; du = 2xdx

$$\begin{aligned}\int_{-\infty}^{\infty} |x| f(x)\,dx &= 2\int_0^{\infty} x\frac{1}{\pi}\left(\frac{1}{1+x^2}\right)dx \\ &= \frac{1}{\pi}\int_0^{\infty}\left(\frac{1}{1+u}\right)du \\ &= [\ln(1+u)]_0^{\infty} \\ &= \infty\end{aligned}$$

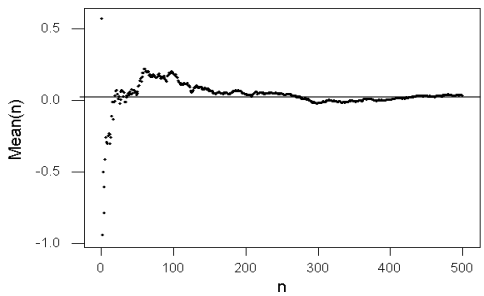
Example 4.1-9: Cauchy Density (continued)

Simulate 500 Cauchy outcomes using Minitab; compute the mean for each n from 1 to 500: Is the mean converging to a value?



Example 4.1-9: Compare Cauchy to Normal

Simulate 500 N(0,1) outcomes using Minitab; compute the mean for each n from 1 to 500: Is the mean converging to a value?



4.1.1 Expectations of Functions of RVs

Suppose Y = g(X), where X is a random variable with probability function p_X or density f_X. What is E(Y)?

Hard way: Find the probability function or density of Y, then compute E(Y) the usual way.

Easy way: E(Y) = E(g(X)), so find average of g(X) directly-

$$E[g(X)] = \begin{cases} \sum_x g(x)p_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx, & \text{if } X \text{ is continuous} \end{cases}$$

provided E(|g(x)|) is finite.

Examples:

Example 4.1-10. Suppose X is Bernoulli(.7). Find E(3X).

Example 4.1-11. Find E(1/X), where X has density:

$$f(x) = \begin{cases} \frac{3}{2}x^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Examples:

Example 4.1-13. Suppose X is continuous with density f.
Let *g(X)* = *a*, where *a* is a constant. Find E(g(X)).

Example 4.1-14. Suppose X is continuous with density f.
Find E[X - E(X)] = E(X - μ)

Expectations of Functions of Jointly Distributed RVs

Suppose $X_1,\ldots,X_n$ are jointly distributed RVs with PF p or PDF f and $Y = g(X_1,\ldots,X_n)$. Then:

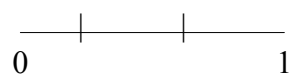
$$E(Y) = \begin{cases} \sum_{x_1,\ldots,x_n} g(x_1,\ldots,x_n)p(x_1,\ldots,x_n), & X_i \text{ discrete} \\ \int\int\cdots\int g(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1,\ldots,dx_n, & X_i \text{ continuous} \end{cases}$$

provided $E(|g(X_1,\ldots,X_n)|)$ is finite.

Corollary: If X and Y are independent random variables and g(X) and h(Y) are fixed functions of X and Y, then:

$$E[g(X)h(Y)] = E[(g(X)] \times E[h(Y)]$$

Example 4.1-15. A stick of length one is broken randomly in two places. What is the average length of the middle piece?



$$\begin{aligned}
E(|U_1 - U_2|) &= \int_0^1\int_0^1 |u_1 - u_2|(1)du_1du_2 \\
&= \int_0^1\int_0^{u_1}(u_1 - u_2)du_2du_1 + \int_0^1\int_{u_1}^1(u_2 - u_1)du_2du_1 \\
&= \int_0^1\left[u_1u_2 - \frac{u_2^2}{2}\right]_0^{u_1} du_1 + \int_0^1\left[\frac{u_2^2}{2} - u_1u_2\right]_{u_1}^1 du_1 \\
&= \int_0^1\left[\frac{u_1^2}{2}\right]du_1 + \int_0^1\left[\frac{1}{2} - u_1 + u_1^2 - \frac{u_1^2}{2}\right]du_1 \\
&= \int_0^1\left[\frac{1}{2} - u_1 + u_1^2\right]du_1 \\
&= \left[\frac{u_1}{2} - \frac{u_1^2}{2} + \frac{u_1^3}{3}\right]_0^1 \\
&= \frac{1}{3}
\end{aligned}$$

4.1.2 Expectations of Linear Combinations of Random Variables

If $X_1,\ldots,X_n$ are jointly distributed RVs with expectations $E(X_i)$ and Y is a linear combination of the X_i:

$$Y = a + \sum_{i=1}^{n} b_i X_i$$

then:

$$E(Y) = a + \sum_{i=1}^{n} b_i E(X_i)$$

Justification: Can show this easily for n = 2 in the continuous case. Assume $f(x_1,x_2)$ gives the bivariate density of X_1 and X_2. Then:

Justification (continued)

To show that the expectation exists, we just use:

$$\begin{aligned}
&\int\int |a + b_1x_1 + b_2x_2| f(x_1,x_2)dx_1dx_2 \\
&\qquad \le \int\int (|a| + |b_1||x_1| + |b_2||x_2|) f(x_1,x_2)dx_1dx_2 \\
&\qquad = |a| + |b_1|E(|X_1|) + |b_2|E(|X_2|) \\
&\qquad < \infty
\end{aligned}$$

Example 4.1-16 Mean of Binomial Distribution

Assume Y ~ Binom(n,p). Hard way to find E(Y):

$$E(Y) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = ???$$

Easy way: Y is the sum of n independent Bernoulli(p) random variables, so:

$$E(Y) = \sum_{k=1}^{n} E(B_i) = \sum_{k=1}^{n} p = np$$

Variance and Standard Deviation

These answer the questions:

- **How spread out is the distribution about its mean?**
- **What is the average squared distance from a random variable to its mean?**

Definition: If X is a random variable with expected value E(X), the variance of X is:

$$Var(X) = \sigma_X^2 = E\{[X - E(X)]^2\}$$

provided the expectation exists. The square root of the variance is the standard deviation, σ_X.

Variance and Standard Deviation

Discrete Case: $Var(X) = \sum_i (x_i - \mu)^2 p(x_i)$

Continuous Case: $Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$

Result: If Var(X) exists and Y = a + bX, then
$Var(Y) = b^2 Var(X)$.

Justification: Since E(Y) = a + bE(X),

Examples

Example 4.2-1: X ~ Bernoulli(p). Find Var(X).

Example 4.2-2: Let X = outcome of a roll of a single die. Find Var(X).

Examples

Example 4.2-3: From example 4.1-17, if X has density:

$$f(x) = \begin{cases} \frac{3}{2}x^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

then E(X) = 0. Find Var(X).

Alternate Formula for the Variance

The variance of X may be calculated as follows:

$$Var(X) = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2$$

Justification:

Examples

Example 4.2-4. X ~ Unif(0,1). Find Var(X).

Example 4.2-5 X ~ N(μ, σ^2). Then E(X - μ)^2 = σ^2. (Text carries out the integration using substitution twice and knowledge of the gamma function (i.e., using some tricks)

Chebyshev's Inequality

Says: The probability that any point is within k standard deviations of the mean <u>for any distribution</u> is at least 1 - 1/k^2 (k need not be integer).

- **k = 2: P(|X - μ| < 2σ) $\geq$ .75**
- **k = 3: P(|X - μ| < 3σ) $\geq$ .8889**
- **k = 4: P(|X - μ| < 4σ) $\geq$ .9375**

Chebyshev's Inequality: Let X be a random variable with mean μ and variance σ^2. Then, for any t > 0,

$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$$

Proof of Chebyshev

Let R = {x: |x - μ| >t}. Then:

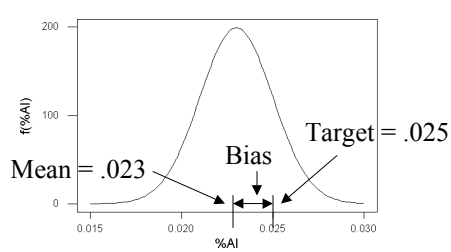
$$
\begin{aligned}
P(|X-\mu|>t) &= \int_R f(x)dx \\
&\le \int_R \frac{(x-\mu)^2}{t^2} f(x)dx \\
&\le \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{t^2} f(x)dx \\
&= \frac{\sigma^2}{t^2}
\end{aligned}
$$

Setting t = kσ, we have:

$$P(|X-\mu|>k\sigma) \le \frac{1}{k^2} \text{ –or– } P(|X-\mu| \le k\sigma) > 1-\frac{1}{k^2}$$

Mean Squared Error (MSE)

- **In describing average distance from a "target," variance does not always give the whole story**
- **Suppose you are manufacturing a drug, and it is important to get the right amount of active ingredient (2.5%) in each tablet. The process distribution of the amount of active ingredient in each tablet and the target are shown below:**



MSE (continued)

Here we want to know the average squared distance $E(X - t)^2$ to the target, t (not to the mean). Recall that:

$$Var(X - t) = E[(X - t)^2] - [E(X - t)]^2$$

Solving for $E(X - t)^2$,

$$\begin{aligned} E[(X-t)^2] &= Var(X-t) + [E(X-t)]^2 \\ &= Var(X) + [E(X) - t]^2 \\ &= \sigma^2 + \beta^2 \end{aligned}$$

or, in English: MSE = Variance + Bias Squared

(Note that MSE = Variance if the bias is zero.)

Applications of MSE

1. Physical measurements (weights, speeds, lengths)

Actual Measurement

= true value + instrument bias + measurement error

$$X = X_t + \beta + \varepsilon$$

$$E(\varepsilon) = 0 \qquad E(X) = X_t + \beta \qquad Var(X) = Var(\varepsilon) = \sigma^2$$

MSE = $\sigma^2 + \beta^2$ = "precision" + "accuracy"

2. Taguchi methods (quality control/quality management)

3. Prediction in statistics (Section 4.4.2, later)

4.3 Covariance and Correlation

Definition: If X and Y are jointly distributed random variables with expectations μ_x and μ_Y, the covariance of X and Y is:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

Notes:

(1) From the definition, Cov(X,X) = Var(X).

(2) Computational form for Cov(X,Y):

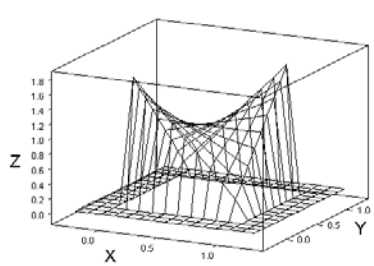
(3) If X and Y are independent,

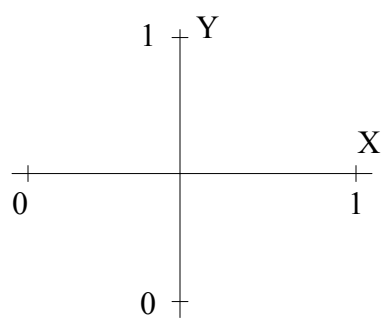
Example 4.3-1:

Recall the joint density from page 126:

f(x,y) = 2 - 2x - 2y + 4xy

on [0,1]x[0,1]; E(X) = E(Y) = .5. From the picture, is the covariance > 0? < 0? 0?





Carrying Out Computation

Covariance of Linear Combinations of RVs

Cov(a + X, Y):

Cov(aX,bY)

Covariance of Linear Combinations of RVs

Cov(X, Y + Z):

Cov(aW + bX,cY + dZ):

Covariance of Linear Combinations of RVs

General case:

$$Cov(\sum_{i=1}^{n} b_i X_i, \sum_{j=1}^{m} d_j Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} b_i d_j Cov(X_i, Y_j)$$

From above it follows that:

$$Var(\sum_{i=1}^{n} b_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j Cov(X_i, X_j)$$

Corollary: If the X_i are independent:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

Example 4.3-2: Variance of a Binomial RV

If X ~ Binomial(n,p), find Var(X).

Correlation Coefficient

Definition: If X and Y are jointly distributed RVs with nonzero variances, then the correlation of X and Y, denoted by ρ, is:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Example 4.3-2. Find the correlation coefficient for the density in Example 4.3-1.

Correlation Theorem

1) $-1 < \rho < 1$

2) $\rho = +1$ if and only if $P(Y = a + bX) = 1$

Proof: See next page for (1), text p 133 for (2).

Implication: ρ is a measure of the strength of the *linear* relationship between X and Y

Proof of Correlation Theorem

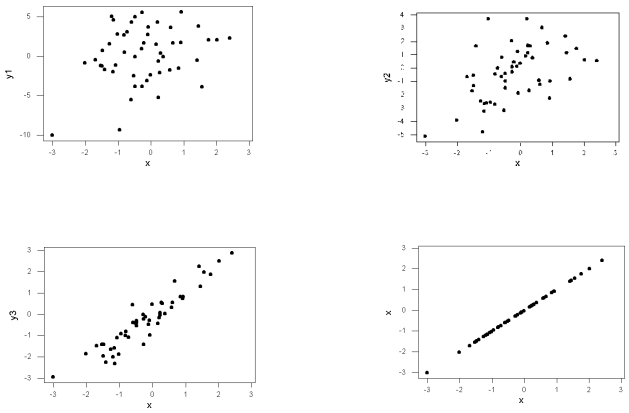
To show $-1 \leq \rho$:

$$\begin{aligned} 0 &\leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2Cov\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{Var(X)}{\sigma_X^2} + \frac{Var(Y)}{\sigma_Y^2} + 2\frac{Cov(X,Y)}{\sigma_X \sigma_Y} \\ &= 2(1+\rho) \end{aligned}$$

To show $\rho \leq 1$:

$$0 \leq Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho)$$

Samples of n = 50 with true correlation = .1, .5, .9, 1.0



4.4 Conditional Expectation and Prediction

Definition: The conditional expectation of Y given X = x is:

$$E(Y|X) = \sum_y y p_{Y|X}(y|x) \qquad \text{discrete case}$$

$$E(Y|X) = \int y f_{Y|X}(y|x)dy \qquad \text{continuous case}$$

Example 4.4-1. Bivariate Normal Distribution.

Recall from Chapter 3, conditional mean of Y given X = x:

$$\mu_{Y|X}(y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

Note: E(Y|X) is a random variable, since X is random!

Conditional Expectation Theorem

Theorem: $E(Y) = E_X[E_Y(Y|X)]$

Justification:

$$
\begin{aligned}
E_X[E(Y|X)] &= \int E(Y|X) f(x) dx \\
&= \int \left[\int y f_{Y|X}(y|x) dy \right] f(x) dx \\
&= \int \int y f_{Y|X}(y|x) f(x) dx dy \\
&= \int y \left[\int f_{XY}(x,y) dx \right] dy \\
&= \int y f_Y(y) dy \\
&= E(Y)
\end{aligned}
$$

Example 4.4-2 Insurance Claims as Random Sums

N = RV giving the number of claims received by an insurance company in one month

X_i = RV giving the amount of the *i*th claim

T = Claims total for the month:

$$T = \sum_{i=1}^{N} X_i$$

Find E(T):

Conditional Variance

Definition. **The conditional variance of Y given X is:**

$$\mathrm{Var}(Y|X) = E(Y^2|X] - [E(Y|X)]^2$$

Note that Var(Y|X) is a random variable, since X is random!

Conditional Variance Theorem:

$$\mathrm{Var}(Y) = \mathrm{Var}_X[E(Y|X)] + E_X\,[\mathrm{Var}(Y|X)]$$

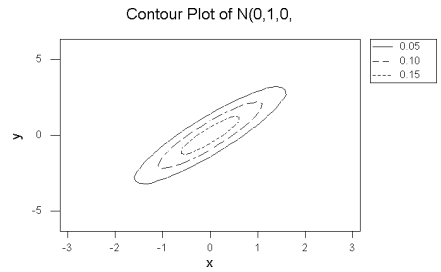
Proof of Conditional Variance Theorem

$$\begin{aligned}
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= E[E(Y^2|X)] - \{E[E(Y|X)]\}^2 \\
&= E[E(Y^2|X)] - E\{[E(Y|X)]^2\} + E\{[E(Y|X)]^2\} - \{E[E(Y|X)]\}^2 \\
&= E[Var(Y|X)] + Var(E[Y|X])
\end{aligned}$$

Example 4.4-3 Find Var(T) for Example 4.4-2.

4.4-2 Prediction

Suppose you are a college admissions officer and you'd like to predict the college GPA (Y) for applicants based on their ACT entrance exam score (X). Suppose also that you know the joint distribution of GPA and ACT scores from past data:



What is the *best* function h(x) for predicting Y, given X?

Criterion: Minimize MSE

Case 1: Suppose we want h(x) to be a constant, c.

What value of c minimizes MSE?

$$\begin{aligned} MSE_{X,Y}[c] &= E_{X,Y}[(Y-c)^2] \\ &= E_{X,Y}[(Y-\mu_Y+\mu_Y-c)^2] \\ &= E_Y[(Y-\mu_Y)^2 + 2(Y-\mu_Y)(\mu_Y-c) + (\mu_Y-c)^2] \\ &= Var(Y) + (\mu_Y-c)^2 = Var + Bias^2 \end{aligned}$$

which is obviously minimized by choosing $c = \mu_Y$

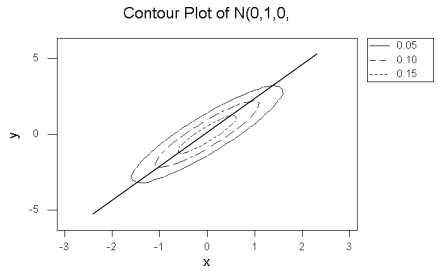
Now onto the real case. Which function h(X) is best?

$$\begin{aligned} MSE_{X,Y}[h(X)] &= E_{X,Y}[(Y-h(X))^2] \\ &= E_X\{E_Y[Y-h(X)|X]^2\} \end{aligned}$$

Best Predictor (continued)

Now at any given X = x, $E[(Y - h(X))^2]$ is minimized by h(x) = E(Y|X = x) as shown in case 1.

Result: The best predictor, given the joint distribution, is the conditional expectation, or the *regression function*.



Best Linear Predictors

In the previous picture, the best function h(x) happened to be linear because X and Y were jointly normally distributed. h(x) is frequently not linear.

If we don't know the joint distribution f(x,y), we generally cannot find h(x). We could ask

- **What is the best linear predictor: $Y = \alpha + \beta X$, or**
- **What is the best quadratic predictor: $Y = \alpha + \beta X + \gamma X^2$**

We will consider the linear case. (Note: this is a parametric model. Many modern nonparametric regression methods, such as Lowess smoothing, attempt to estimate h(x) directly)

Best Linear Predictor

Choose α to minimize term 2:

Choose β to minimize term 1:

Best Linear Predictor (continued)

Result:

$$\begin{aligned}\hat{Y} &= \alpha + \beta x \\ &= (\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X) + \rho\frac{\sigma_Y}{\sigma_X}X \\ &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\end{aligned}$$

For this optimal predictor,

$$\begin{aligned}MSE &= E[(Y - \hat{Y})^2] \\ &=\end{aligned}$$

Advantage for Linear Predictor:

Only requires knowledge of μ_X, μ_Y, σ_X, σ_Y, and σ_{XY}, not the entire joint distribution.

We can usually estimate these using sample means, variances and covariances. This leads to the least squares regression line.

4.5 The Moment Generating Function (MGF)

A particularly useful trick for determining the expected value of a sum of random variables (without having to endure the pain of deriving the "convoluted" density.)

***Definition*: The moment-generating function of the random-variable X is given by:**

$$
\begin{aligned}
M_X(t) &= E[e^{tX}] \\
&= \sum_{i=1}^{n} e^{tx_i} p(x_i) && \text{if } X \text{ is discrete} \\
&= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx && \text{if } X \text{ is continuous}
\end{aligned}
$$

Using the MGF to find Moments

Theorem: Let X be a random variable with MGF $M_X(t)$. Then the *r*th non-central moment of X is:

$$E[X^r] = \mu'_r = \left[\frac{d^r M_X(t)}{dt^r}\right]_{t=0} = M_X^{(r)}(0)$$

Proof:

$$\frac{d^r M_X(t)}{dt^r} = \sum_{i=1}^{n} x_i^r e^{tx_i} p(x_i) \qquad \textbf{(discrete)}$$

$$= \int_{-\infty}^{\infty} x^r e^{tx} f(x)dx \qquad \textbf{(continuous)}$$

Setting t = 0, both cases yield $E[X^r]$.

Example 4.5-1 Binomial MGF

Find $M_X(t)$ if X ~ Binomial(n,p), and use it to find the mean and variance of X.

Example 4.5-2 Poisson MGF

Find $M_X(t)$ if $X \sim \text{Poisson}(\lambda)$, and use it to find the mean and variance of X.

Example 4.5-2 Standard Normal MGF

Find $M_X(t)$ if $X \sim N(0,1)$, and use it to find the mean and variance of X.

Other Properties of the MGF

***Uniqueness Theorem*. Let X and Y be two random variables with MGFs $M_X(t)$ and $M_Y(t)$, respectively. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the same probability distribution.**

$$M_X(t) = M_Y(t) \Leftrightarrow F_X = F_Y$$

Addition of a Constant:

$$M_{X+a}(t) = e^{at} M_X(t)$$

Multiplication by a Constant

$$M_{aX}(t) = M_X(at)$$

Example 4.5-3 General Normal Distribution

Let X be a standard normal random variable and let:

$$Y = \mu + \sigma X$$

then Y ~ N(μ,σ^2). Find the MGF of Y:

Other Properties of the MGF (Continued)

***Sums of Independent RVs*. Let X_1 and X_2 be independent random variables with MGFs $M_{X1}(t)$ and $M_{X2}(t)$, respectively, and $Y = X_1 + X_2$.**

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t)$$

Justification:

Example 4.5-4 Sum of Independent Normals

If $X \sim N(\mu,\sigma^2)$ and, independent of X, $Y \sim N(\nu,\tau^2)$, then the MGF of X + Y is:

$$e^{\mu t}e^{t^2\sigma^2/2}e^{\nu t}e^{t^2\tau^2/2} = e^{(\mu+\nu)t}e^{t^2(\sigma^2+\tau^2)/2}$$

which is the MGF of a normal distribution with mean $\mu + \nu$, and variance $\sigma^2 + \tau^2$.

Sums of independent normals are normal!

4.6 Approximate Methods in Statistics

- **Suppose we know the mean and variance of X**
- **We want to know the mean and variance of Y = g(X) where g(X) is a nonlinear function of X.**

Example: X has mean μ, and variance σ^2. Distribution is unknown. Can we find (or approximate) the mean and variance of 1/X? Of ln(X)? Of X^2?

Note: If g(X) were a linear function, Y = aX + b,

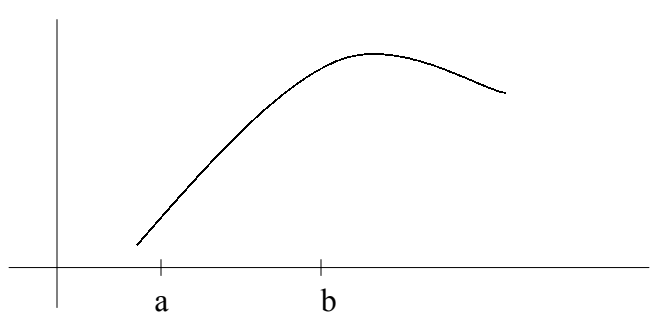
E(Y) =

Var(Y) =

Trick:
Use a Linear or Quadratic Approximation of g

Taylor Series representation of a function:

$$g(b) = g(a) + (b-a)g'(a) + \frac{(b-a)^2}{2!}g^{(2)}(a) + \frac{(b-a)^3}{3!}g^{(3)}(a) + \cdots$$



To approximate g, just use the first few terms; drop rest

Example: 4.6-1

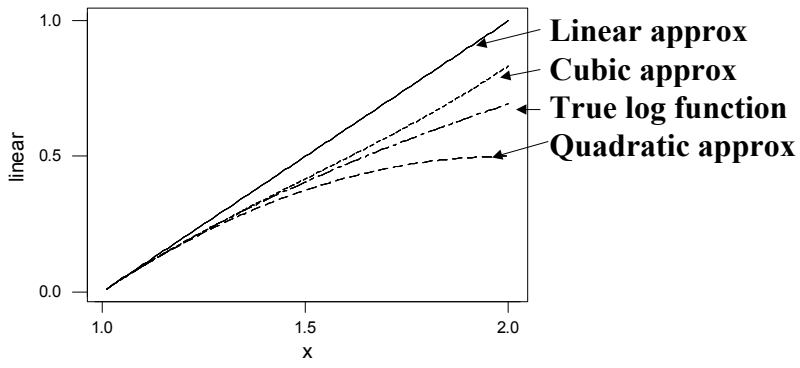
First, second, and third-order Taylor approximations to g(X) = log(X) in the neighborhood of X = 1:

First order:

Second order:

Third order:

Picture



Start with a first-order expansion to obtain an approximate expression for E(g(X)) and Var(g(X)):

$$g(b) = g(a) + (b-a)g'(a)$$

This approximation will be good if g is nearly linear in a range in which X has high probability.

Expanding around the mean, use $a = \mu_X$ and $b = X$:

$$g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X)$$

Then E(g(X)) =

Var(g(X)) =

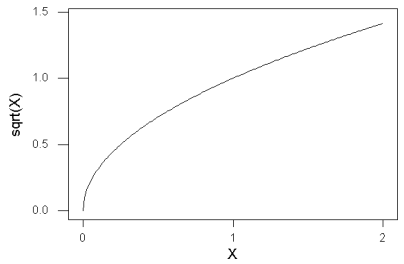
Can Improve E(g(X)) with a second-order expansion:

$$g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{(X-\mu_X)^2}{2!}g^{(2)}(\mu_X)$$

Then:

E[g(X)] =

Example 4.6-2



(a) Suppose X ~ Unif[0,1], and Y =g(X) = SQRT(X). Give approximate values of E(Y) and Var(Y) using the delta method; compare to the exact values to assess the accuracy of the approximation

Example 4.6-2a (continued)

Example 4.6-2b

b) Rework part (a), this time with X ~ Unif (1, 2). (Before you do any calculations, do you think the approximation improve or deteriorate? Why?)

Chapter 5: Limit Theorems

Three essential concepts:

1) The Law of Large Numbers (LLN)
2) Convergence in Distribution
3) The Central Limit Theorem

All concern the limiting distribution of a sum of independent random variables.

The Law of Large Numbers concerns the limiting value of the an average, the central limit theorem concerns the distribution of sums (or averages) of independent RVs

5.2 The Law of Large Numbers

Example 5.2-1. Flip a coin n times. $X_i = 1$ if the *i*th flip is a head, and 0 if tail. $X_i \sim$ Bernoulli (.5).

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \text{ sample probability after n flips}$$

How do we know that

$$p_n \to p \text{ as } n \to \infty?$$

More generally, how do we know that

$$\bar{X}_n \to E(X) \text{ as } n \to \infty?$$

Convergence in Probability and the LLNs

Definition. A sequence of random variables Z_n is said to converge in probability to α if, for any $\varepsilon > 0$,

$$P(|Z_n - \alpha| > \epsilon) \to 0, \text{ as } n \to \infty$$

Theorem. (Weak) Law of Large Numbers.

Let $X_1, X_2, \ldots, X_i, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X}_n = \Sigma X_i/n$. Then, for any $\varepsilon > 0$,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0, \text{ as } n \to \infty$$

In other words, $\bar{X}_n$ converges in probability to μ.

Proof of LLNs

(1) $E(\bar{X}_n) =$

(2) $Var(\bar{X}_n) =$

(3) Chebyshev's inequality tells us that for any $\varepsilon > 0$:

Example 5.2-2 Monte Carlo Integration

Suppose we wish to calculate:

$$I(g) = \int_0^1 g(x)dx$$

but g is too complicated to integrate analytically. Here's how:

(1) Generate $X_1,\ldots,X_n$ as independent Uniform[0,1] random variables; density of X_i is $f_X(x) = 1$, $0 < X_i < 1$.

(2) Compute:

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

(3) By LLNs,

$$\bar{g}_n \to E[g(X)] = \int_0^1 g(x) f_X(x)dx = \int_0^1 g(x)dx = I(g)$$

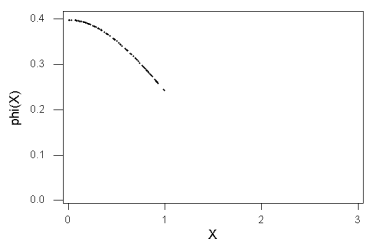
(4) For large n, $\bar{g}_n$ will be close to I(g)!!!

Example 5.2-3

With the help of Minitab, use Monte-Carlo integration and n = 100 to approximate:

$$\int_0^1 \phi(x)dx = \frac{1}{2\pi}\int_0^1 e^{\frac{x^2}{2}}dx$$

The sample average is .337; true value Φ(1) - Φ(0) = .3413



5.3 Convergence in Distribution

Definition. Let $X_1, X_2, \ldots, X_i, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$, and let X be a random variable with CDF F. We say that X_n *converges in distribution* to X if:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every point at which F is continuous.

Continuity Theorem. Let F_n be a sequence of CDFs with the corresponding MGF M_n. Let F be a CDF with the MGF F. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F.

Example 5.3-1

Using the continuity theorem, show that as the Poisson parameter $\lambda_n \to \infty$, the Poisson distribution converges to the normal distribution with mean λ_n and variance λ_n.

Let $\lambda_1, \lambda_2, \ldots$ be an increasing sequence of Poisson parameters ($\lambda_n \to \infty$) and $X_i \sim \text{Pois}(\lambda_i)$. Let:

$$Z_n = \frac{X_n - E(X_n)}{\sqrt{Var(X_n)}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$$
$$= \frac{1}{\sqrt{\lambda_n}} X_n - \sqrt{\lambda_n}$$

We know that $E(Z_n) = 0$, $Var(Z_n) = 1$.

Example 5.3-1 (continued)

We need to show:

$$M_{Z_n}(t) \to M_Z(t) = e^{\frac{t^2}{2}}$$

Since $X_n \sim \text{Pois}(\lambda_n)$,

$$M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$$

Therefore:

$$M_{Z_n}(t) = e^{-\sqrt{\lambda_n}t} M_{X_n}\left(\frac{t}{\sqrt{\lambda_n}}\right)$$
$$= e^{-\sqrt{\lambda_n}t} e^{\lambda_n(e^{\frac{t}{\sqrt{\lambda_n}}} - 1)}$$

Example 5.3-1 (continued)

Taking logs to make life easier:

$$log(M_{Z_n}(t)) = -\sqrt{\lambda_n}t + \lambda_n \left[\sum_{k=0}^{\infty} \frac{\left(\frac{t}{\sqrt{\lambda_n}}\right)^k}{k!} - 1 \right]$$

$$= -\sqrt{\lambda_n}t - \lambda_n + \frac{\lambda_n \left(\frac{t}{\sqrt{\lambda_n}}\right)^0}{0!} + \frac{\lambda_n \left(\frac{t}{\sqrt{\lambda_n}}\right)^1}{1!} + \frac{t^2}{2} + \sum_{k=3}^{\infty} \frac{\frac{t^k}{\lambda_n^{k/2-1}}}{k!}$$

which $\rightarrow t^2/2$, as $\lambda_n \rightarrow \infty$, which is what we needed to show.

Example 5.3-2 Normal Approximation to Poisson

On average, 1200 cars per hour cross the Interstate 94 Mississippi River bridge during the evening. What is the probability that more than 1300 cars cross during a given evening hour?

Central Limit Theorem

Let $X_1, X_2, \ldots, X_i, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2 and the common CDF F and MGF M defined in a neighborhood of zero. Let:

$$S_n = \sum_{i=1}^{n} X_i$$

Then:

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \quad -\infty < x < \infty$$

"Sums of Independent Random Variables are Approximately Normal"

Notes on Statement of Central Limit Theorem

Let

$$S_n = \sum_{i=1}^{n} X_i$$

$$\sigma_{S_n}^2 = Var(S_n) = Var(\sum X_i) = n\sigma^2$$

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu_{S_n}}{\sigma_{S_n}} \le x\right) = \lim_{n\to\infty} P\left(\frac{S_n - 0}{\sigma\sqrt{n}} \le x\right)$$
$$= P(Z \le x) = \Phi(x)$$

Proof: The proof consists of showing that the MGF of the standardized sum converges to the MGF of a standard normal. We need a couple of preliminary "fun facts" before we do the proof.

More Fun Facts

1) **Suppose:**

$$\lim_{n\to\infty} g_n(t) = g(t)$$

Then:

$$\lim_{n\to\infty}\left(1+\frac{g_n(t)}{n}\right)^n = e^{g(t)}$$

2) **Consider the second-order Taylor expansion:**

$$f(s) = f(0) + sf'(s) + \frac{s^2}{2}f''(s) + \epsilon_s$$

where:

$$\epsilon_s = \frac{s^3}{3!}f^{(3)}(s) + \frac{s^4}{4!}f^{(4)}(s) + \cdots$$

Fact:

$$\frac{\epsilon_s}{s^2} \to 0 \text{ as } s \to 0$$

Proof of Central Limit Theorem (page 1)

Let:

$$M_{S_n}(t) = [M_X(t)]^n$$

Then:

$$M_{Z_n}(t) = \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

We "simply" need to show that:

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{t^2}{2}}$$

Proof of Central Limit Theorem (page 2)

Amazing Trick: **We don't know $M_X(t)$, because we want to prove this result for *any* distribution, therefore for any MGF $M_X(t)$. We'll use a second-order Taylor approximation to $M_X(t)$ that will be good for *any* $M_X(t)$, expanding around $E(X_i) = \mu_X = 0$:**

$$M_X(s) = M_X(0) + sM_X'(0) + \frac{s^2}{2}M_X''(0) + \epsilon_s$$

We can get away with this because we know:

$$\begin{aligned} M_X'(0) &= E(X_i) = 0 \\ M_X''(0) &= E(X_i^2) = Var(X_i) + [E(X_i)]^2 \\ &= Var(X_i) = \sigma^2 \end{aligned}$$

and so:

$$M_X(s) = 1 + \frac{s^2}{2}\sigma^2 + \epsilon_s$$

Proof of Central Limit Theorem (page 3)

Then, with $s = t/(\sigma\sqrt{n})$

$$\begin{aligned} M_{Z_n}(t) &= \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left[1 + \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2}\sigma^2 + \epsilon_n\right]^n \\ &= \left[1 + \frac{t^2}{2n} + \epsilon_n\right]^n \end{aligned}$$

Now by FF2:

$$\begin{aligned} \frac{\epsilon_n}{s^2} &= \left(\frac{\sigma^2}{t^2}\right) n\epsilon_n \\ &= \text{constant} \times n\epsilon_n \to 0 \text{ as n } \to \infty \end{aligned}$$

Proof of Central Limit Theorem (page 4)

Therefore:

$$n\epsilon_n \to 0 \text{ as n } \to \infty$$

So:

$$M_{Z_n}(t) = \left[1 + \frac{\frac{t^2}{2} + n\epsilon_n}{n}\right]^n$$
$$= \left[1 + \frac{g_n(t)}{n}\right]^n$$

Then by FF1, since:

$$\lim_{n\to\infty} g_n(t) = \frac{t^2}{2}$$
$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{t^2}{2}}$$

OK, So What?

We know that sums of independent, identically distributed random variables are asymptotically normally distributed.

<u>The real question:</u> **How large does n have to be before we can safely assume the sum is "approximately" normal?**

In the following examples, we will use the Minitab "execs" on the next pages to simulate the distribution of sums of iid random variables for varying distributions and varying *n*.

We will use samples of size n = 1, 5, 20, and 50, and we will generate 500 random sums in each case.

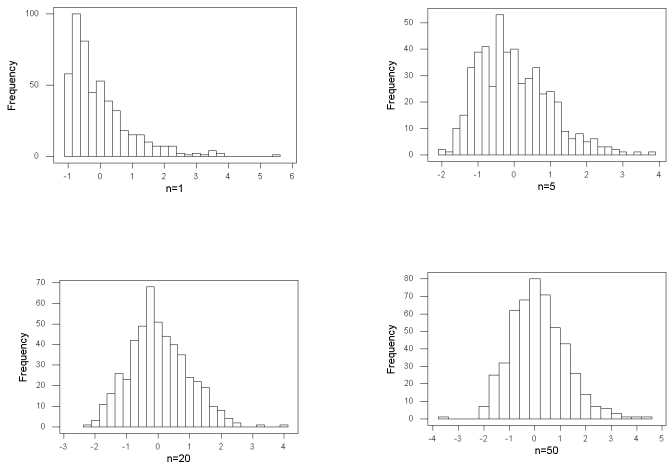
VARYN.TXT

```
let k3 = 0
let k1 = 1
let k2 = 2
exec 'c:\files\teaching\6850--Math Stats\randomsum.txt' 500
let k3 = 0
let k1 = 5
let k2 = 3
exec 'c:\files\teaching\6850--Math Stats\randomsum.txt' 500
let k3 = 0
let k1 = 20
let k2 = 4
exec 'c:\files\teaching\6850--Math Stats\randomsum.txt' 500
let k3 = 0
let k1 = 50
let k2 = 5
exec 'c:\files\teaching\6850--Math Stats\randomsum.txt' 500
```

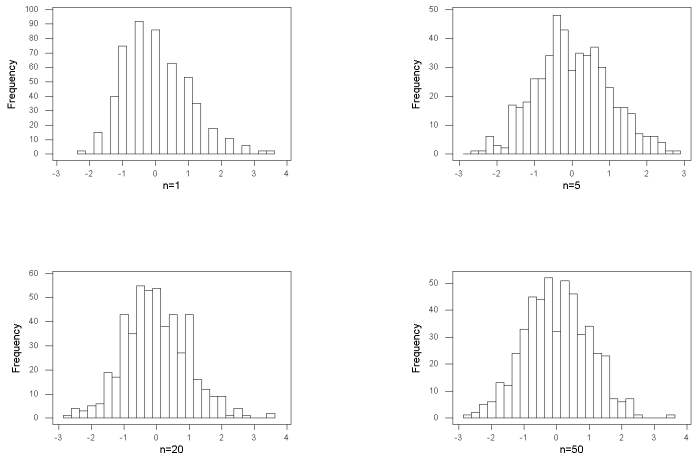

RANDOMSUM.TXT

```
let k3 = k3 + 1
random k1 c1;
  exponential .2.
let ck2(k3) = sum(c1 -(1/5))/((1/5)*sqrt(k1))
```

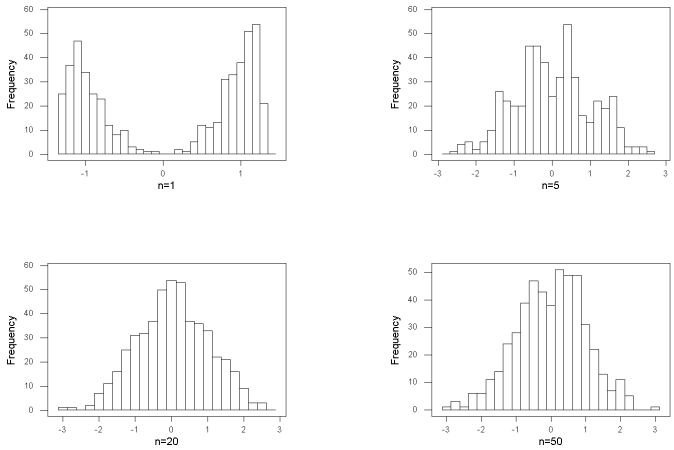
Example 5.3-1 X ~ Exponential(5)



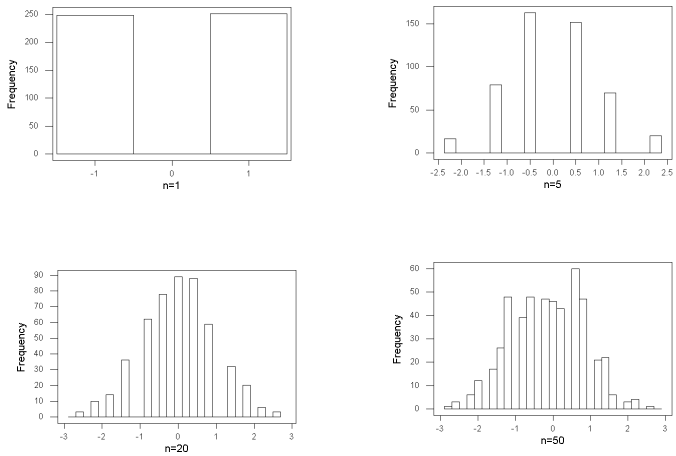
Example 5.3-2 X ~ Poisson(5)



Example 5.3-3 $f(x) = 1.5\,x^2, \quad -1 \le x \le 1$ (bathtub-shaped distribution)



Example 5.3-4 X ~ Bernoulli (.5)



Normal Approximation to the Binomial

If Xn ~ Binomial(n,p), then:

$$X_n = \sum_{i=1}^{n} B_i$$

where B_i is a Bernoulli(p) random variable. From the central limit theorem:

$$\frac{X_n - \mu_{X_n}}{\sigma_{X_n}} = \frac{X_n - np}{np(1-p)}$$

converges to a standard normal random variable. So for large n,

$$P(X_n \le x) = P\left(Z \le \frac{x - np}{np(1-p)}\right)$$

Normal Approximation to the Binomial

Example 5.3-5 X ~ Binomial(30, .3) Find P(X ≤ 11), using an appropriate normal approximation, and compare to the exact value.

Chapter 6: Distributions Derived from the Normal

The χ^2, t, and F distributions are central to data analysis:

χ^2 **Relates to the distribution of the sample variance--allows inference about σ^2. Central in the analysis of categorical data (log-linear models, logistic regression) and multivariate statistics (factor analysis)**

$$s^2 \text{ is distributed as } \frac{\sigma^2\chi^2}{n-1}$$

t: **Central to the analysis of sample means, regression analysis; t is just an "estimated Z value":**

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad \hat{Z} = t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$$

F: **ANOVA and regression; testing differences in variances**

Definitions

Definition 1. **If $Z \sim N(0,1)$, then $U = Z^2 \sim \chi^2_1$ (i.e., Z^2 has a chi-square distribution with one degree of freedom).**

Definition 2. **If $Z_1, Z_2, \ldots, Z_n$ are *iid* N(0,1), then $U = \Sigma Z_i^2 \sim \chi^2_n$.**

Definition 3. **If $Z \sim N(0,1)$, $U \sim \chi^2_n$, and U and Z are independent, then**

$$t = \frac{Z}{\sqrt{\frac{U}{n}}}$$

follows a t distribution with n degrees of freedom

Definitions (continued)

***Definition 4*. If U ~ χ^2_m, V ~ χ^2_n, and U and V are independent, then:**

$$F = \frac{U/m}{V/n}$$

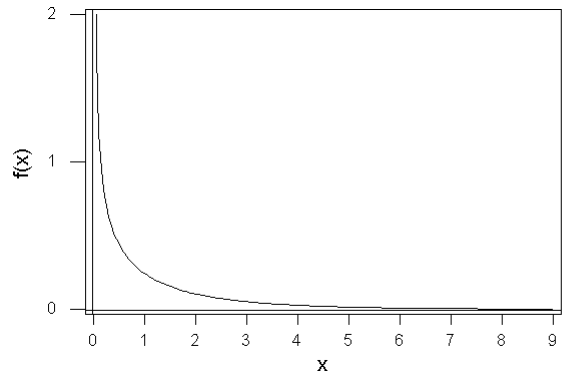
follows an F distribution with m numerator and n denominator degrees of freedom.

1. Derivation of Density of U = Z^2

We did this before (see page 109). Recall:

We also noted that this is a gamma(.5, .5) density.

Picture of chi-square (1) density



2. Derivation of $\chi^2(n)$: Density of $U = \Sigma Z_i^2$

***Preliminary Result*: If $X_i \sim$ gamma(α_i, λ) for i= 1,…,n and the X_i's are independent, then the sum:**

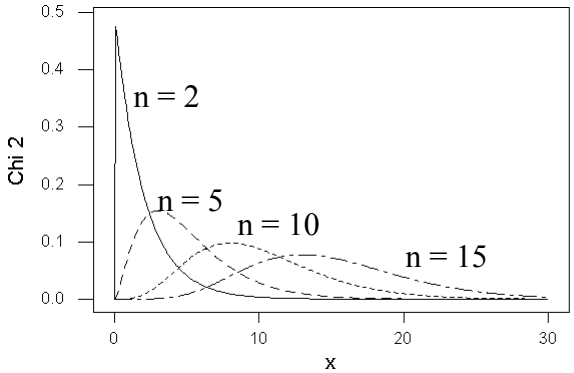
$$Y = X_1 + X_2 + \ldots + X_n \sim \text{gamma}(\alpha_1 + \alpha_2 + \ldots + \alpha_n, \lambda)$$

Sums of gammas having the same scale parameter are gamma!

***Proof:* We'll first find the MGF of a gamma random variable, and then find the MGF of the sum of gammas. It will then follow that U ~ gamma(n/2,.5)**

Derivation of $\chi^2(n)$ Density (page 2)

Some $\chi^2(n)$ Densities



Example 6.2-1

If U ~ χ^2_n, find E(U).

3. Derivation of t Density

The t density is given by:

$$f(t) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \qquad -\infty < t < \infty$$

Proof:

$$t = \frac{Z}{\sqrt{\frac{U}{n}}}$$

where Z is N(0,1) and U is χ^2_n. Since U and Z are independent, their joint density is:

$$f(z,u) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\frac{1}{2^{n/2}\Gamma(n/2)}u^{n/2-1}e^{-u/2}$$

3. Derivation of t Density (continued)

We will define a second random variable V = U. Then we have:

$$t = z/\sqrt{u/n} \qquad v = u$$

Solving for Z and U, we get:

$$z = t\sqrt{v}/\sqrt{n} \qquad u = v$$

from which we obtain:

$$J = \begin{vmatrix} \sqrt{v/n} & t/2\sqrt{vn} \\ 0 & 1 \end{vmatrix} = \frac{\sqrt{v}}{\sqrt{n}}$$

Therefore:

$$g(t,v) = \frac{1}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)} v^{n/2-1} e^{-\{(v/2)[1+(t^2/n)]\}} \frac{\sqrt{v}}{\sqrt{n}}$$

3. Derivation of t Density (continued)

Integrating out v gives the marginal distribution of t (and the result). The integration requires substitution using:

$$w = \frac{v(1+\frac{t^2}{n})}{2} \text{ and } dv = \frac{2dw}{(1+\frac{t^2}{n})}$$

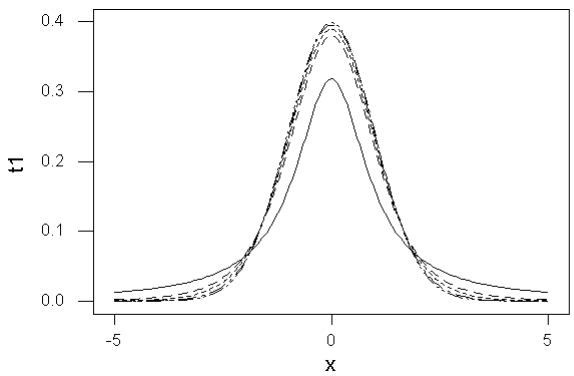
and some tedious algebra.

Notes on t:

- **Developed in 1908 by W. S. Gosset**
- **Worked at Guinness Brewery, which did not permit employees to publish**
- **So Gosset published under the name "Student"**
- **Still referred to as "Student's" t distribution**

Some t distributions

df = 1, 5, 10, 30, ∞



4. Derivation of the F Distribution

Derivation is basically the same as that for the t distribution:

1) $$F = \frac{\chi^2_m/m}{\chi^2_n/n} = \frac{U/m}{V/n}$$

2) The joint density of U and V is the product of the individual chi-square densities, by independence.

3) Find the joint density of F = (U/m)/(V/n) and W = V in the usual way (change of variable formula).

4) Integrate over w to find the marginal distribution of F

6.3 The Sample Mean and the Sample Variance

We will assume that $X_1, X_2, \ldots, X_n$, are iid $N(\mu,\sigma^2)$. How are the sample mean and variance:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

distributed?

Distribution of the Sample Mean

Theorem 1: $\bar{X} \sim N(\mu,\sigma^2/n)$.

Proof:

Distribution of the Sample Variance

Theorem 2:

$$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2_{n-1}$$

Proof:

Distribution of Sample Variance (continued)

Distribution of the t-score

Theorem 3: s^2 and $\bar{X}$ are independent.

Proof: Difficult--beyond this course and omitted.

Theorem 4:

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx t_{n-1}$$

Proof:

Distribution of the F-ratio

Theorem 5: Suppose that $X_1, X_2, \ldots, X_{n1}$, are iid $N(\mu_1, \sigma^2)$. Suppose also that, $Y_1, Y_2, \ldots, Y_{n2}$, are iid $N(\mu_2, \sigma^2)$ and independent of $X_1, X_2, \ldots, X_{n1}$. Let s_1^2 be the sample variance of $X_1, X_2, \ldots, X_{n1}$ and s_2^2 be the sample variance of $Y_1, Y_2, \ldots, Y_{n2}$. Then the ratio of the sample variances:

$$F^* = \frac{s_1^2}{s_2^2}$$

has an F distribution with n_1-1 and n_2-1 df. (Note that the variance σ^2 is the same for both samples; μ_i's need not be!)

Proof:

Examples

Example 6.3-1 Suppose that $X_1, X_2, \ldots, X_{25}$ are an iid sample from N(100, 16) with sample mean $\bar{X}$ and sample variance s_1^2.

a) Superimpose plots of the distribution of X_i and the distribution of $\bar{X}$.

b) Find $P(X > 98)$ and $P(\bar{X} > 98)$

Example 6.3-1 (continued)

c) Sketch and find $P(s_1^2 > 30)$.

d) Suppose that $Y_1, Y_2, \ldots, Y_{36}$ are a second iid sample from N(100, 16) with sample mean $\bar{Y}$ and sample variance s_2^2. Sketch and find $P(s_1^2 / s_2^2 > 2)$.

Chapter 7: Survey Sampling

Introduces notions of:

- **Simple random sampling (s.r.s.)**
- **Sampling distributions**
- **Sample statistics: mean, variance, total, proportion**
- **Sampling in finite populations**
- **Confidence intervals**
- **Stratified random sampling**

Note: We're beginning to move into applied statistics!

7.2 Population Parameters

We will consider two types of populations:

1. **Finite population of size N: $\{x_1, x_2, \ldots, x_N\}$, and parameters:**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \tau = \sum_{i=1}^{N} x_i = N\mu$$

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

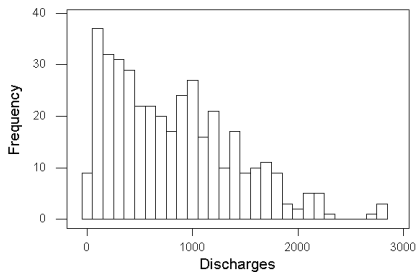
2. **An infinite population or an infinite process: $\{x_1, x_2, x_3, \ldots\}$**

$$\mu = \int_{-\infty}^{\infty} x f(x)dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Finite Population Example (7.2-1)

The population consists of N = 393 short-stay hospitals, and the measurement of interest, x_i, is the number of patients discharged during January 2000. The histogram below summarizes the population



Finite Population Example (7.2-1)

Population parameters are:

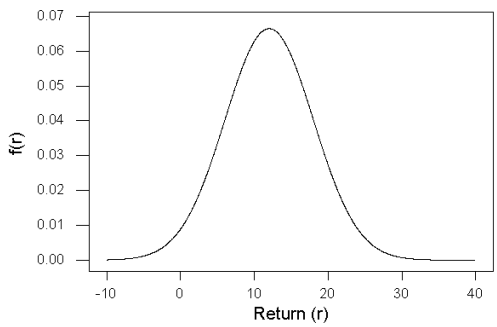
```
MTB > let k2 = (1/393)*sum((discharges - k1)**2)
MTB > let k3 = sqrt(k2)
MTB > let k4 = sum(discharges)
MTB > print k1-k4
```

Data Display

```
K1    814.603
K2    346882
K3    588.966
K4    320139
```

Infinite Population or Process Example (7.2-2)

Annual returns on stocks in a particular industry are believed to follow a normal distribution with mean 12% and standard deviation 6%.



Notes on Finite Population Parameters

Note 1:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \\ &= \frac{1}{N}\left(\sum_{i=1}^{N}x_i^2 - 2\mu\sum_{i=1}^{N}x_i + N\mu^2\right) \\ &= \frac{1}{N}\left(\sum_{i=1}^{N}x_i^2 - 2N\mu^2 + N\mu^2\right) \\ &= \frac{1}{N}\left(\sum_{i=1}^{N}x_i^2\right) - \mu^2\end{aligned}$$

Note 2. If x_i is a zero-one (Bernoulli) variable for all i, then:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{N}x_i^2\right) - \mu^2 \\ &= \frac{1}{N}\left(\sum_{i=1}^{N}x_i\right) - \mu^2 \ \text{(since } x_i = x_i^2\text{)} \\ &= p - p^2 = p(1-p)\end{aligned}$$

Simple Random Sampling (s.r.s.)

Simple random samples are constructed using a probability mechanism that assures:

1) ***Each element of the population has an equal probability of being selected.***
2) ***Each possible sample of size n has the same probability of occurrence.***

s.r.s Notes

1. Number of possible samples $= \binom{N}{n}$

2. $P(\text{any one sample}) = \dfrac{1}{\binom{N}{n}}$

3. $P(x_i \text{ is selected}) =$

4. $$\begin{aligned} P(X_i = x_j) &= P(x_j \text{ is selected AND } x_j \text{ occupies ith place in sample}) \\ &= P(x_j \text{ is selected}) P(X_i = x_j | x_j \text{ is selected}) \\ &= \frac{n}{N}\frac{1}{n} = \frac{1}{N} \end{aligned}$$

7.3.1 Expectation and Variance of the Sample Mean and Total

For finite population sampling, s.r.s induces a probability distribution on the sample elements $X_1, X_2, \ldots, X_n$.

Find: $E(X_i)$, $E(\bar{X})$, and $E(T)$, where T is the sample total

Are $\bar{X}$ and T unbiased?

Are X_i and X_j independent for $i \neq j$?

Find $Cov(X_i, X_j)$

$Cov(X_i, X_j)$ (continued)

Find $Var(\bar{X})$, $Var(T)$, $Var(\hat{p})$

Summary

		Population Variances of Statistic	
Parameter	**Statistic**	**Finite Pop**	**Infinite Pop**

Example 7.3-1

a) **Compute the true variance of the average of a sample of size 32 drawn from the population of short-stay hospitals; recall N = 393, μ = 814.603, and σ^2 = 346,882.**

b) **Using Minitab, draw a sample of size 32 from the population of short-stay hospitals; estimate the mean.**

Example 7.3-1 (continued)

c) **In the population of short-stay hospitals, let B_i be an indicator variable that is 1 if the ith hospital had fewer than 1000 discharges, and 0 otherwise. (p = .654). Compute the standard deviation of the sample proportion if n = 32.**

Estimating σ^2

Since $E[X - \mu_X]^2 = N^{-1}\Sigma(X_i - \mu)^2$ for s.r.s, it's natural to think of using:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

Theorem: With s.r.s,

$$E[\hat{\sigma}^2] = \sigma^2 \left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right)$$

Note 1: Apparently, $\hat{\sigma}^2$ will be a little to small, on average. We say $\hat{\sigma}^2$ will is a *biased estimator* of σ^2 since its expected value is not equal to σ^2.

Note 2: Easily fixed. How?

Proof of Theorem

Unbiased Estimators

1. An unbiased estimator of σ^2 is:

$$\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\hat{\sigma}^2 = s^2\left[\frac{N-1}{N}\right] \quad \text{for finite population}$$

$$= s^2 \quad \text{for infinite population}$$

where:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

2. An unbiased estimator of Var($\bar{X}$) is:

Unbiased Estimators

3. An unbiased estimator of Var($\hat{p}$) is:

4. An unbiased estimator of Var(T) is:

Examples

Example 7.3-2: Draw a sample of size 32 from the population of short-stay hospitals; estimate the variance, and the variance of the $\bar{X}_{32}$.

Example 7.3-3. Compute $\sigma_{\hat{p}}$; then estimate $\hat{p}$ and the variance of p based on a sample of size 32. (p is the proportion of hospitals that had fewer than 1000 discharges--see example 7.3-1c.)

7.3.3 Normal Approximation to Sampling Distribution of $\bar{X}$

Recall the Central Limit Theorem: As n $\to \infty$,

$$P\left(\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \le z\right) \to \Phi(z)$$

We have two problems applying it in finite populations:

(1)

(2)

But…

(1) There are other central limit theorems that work in the finite population sampling context, and

(2) As long as n is "large" the approximation is good

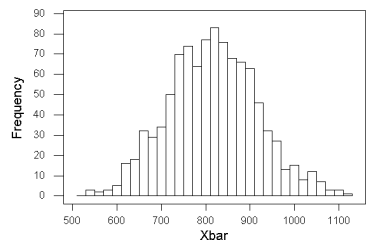
Example 7.3-3

Draw 1000 samples of size 32 from the population of short-stay hospitals, computing and storing $\bar{X}_{32}$ each time. Draw a histogram of $\bar{X}_{32}$ and compare the variance of the simulated population of sample averages to the true variance of $\bar{X}_{32}$.

Recall: Var($\bar{X}_{32}$) =

Results

Distribution of $\bar{X}_{32}$:




```
MTB > let k1 = mean(c4)
MTB > let k2 = sum((c4-k1)**2)/1000
MTB > print k1 k2
K1    815.868
K2    10043.6
```


Summary

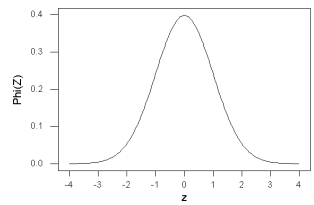
Parameter	Estimate	Standard Error	Estimated Std Error
μ	$\bar{X}$	$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]}$	$s_{\bar{X}} = \sqrt{\frac{s^2}{n}\left[1-\frac{n}{N}\right]}$
τ	$T = N\bar{X}$	$N\sigma_{\bar{X}} = N\sqrt{\frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]}$	$Ns_{\bar{X}} = N\sqrt{\frac{s^2}{n}\left[1-\frac{n}{N}\right]}$
p	$\hat{p}$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}\left[\frac{N-n}{N-1}\right]}$	$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left[1-\frac{n}{N}\right]}$

Confidence Intervals for μ

By the Central Limit Theorem,

$$\bar{X} \sim N(\mu, \sigma_{\bar{x}})$$

for n and N large enough. So we have the following sampling distribution for $Z = (\bar{X} - \mu)/\sigma_{\bar{X}}$



Derivation of CI for μ

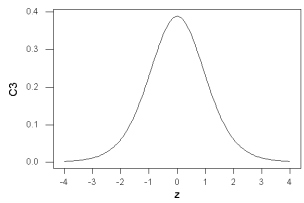
Since:

$$P(-z(\alpha/2) \leq Z \leq z(\alpha/2)) = 1 - \alpha$$

For σ unknown and $X_i \sim N(\mu, \sigma^2)$

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \sim t \qquad \text{with n-1 df}$$

Picture:



Then:

Summary: Some Confidence Intervals

CI for:	σ known	σ unknown
μ	$\bar{X} \pm \sigma_{\bar{X}} z(\alpha/2)$	$\bar{X} \pm s_{\bar{X}} t(\alpha/2)$
τ	$N\bar{X} \pm N\sigma_{\bar{X}} z(\alpha/2)$	$N\bar{X} \pm N s_{\bar{X}} t(\alpha/2)$
p	$\hat{p} \pm \sigma_{\hat{p}} z(\alpha/2)$	$\hat{p} \pm s_{\hat{p}} z(\alpha/2)$

For use of the t distribution, the X_is (observations) should be approximately normally distributed.

Examples

Example 7.3-4. A particular area contains 8000 condominium units. In a survey, occupants were asked how many motor vehicles they owned. A random sample of size 100 yields $\bar{X} = 1.6$, and $s = .8$. Give a 95% confidence interval for μ and for τ.

Examples

Example 7.3-5. In the same survey, 12% said they planned to sell their condominiums within the next year. Give a 90% confidence interval for p, the true proportion who will sell in the next year.

Sample Size Planning

In example 7.3-4, the "half-width" of the interval turned out to be .16. How large a sample will be required if the desired half-width of the interval is H = .10? Ignore the finite population correction (fpc), and assume that σ is known to be .8.

Summary: Sample Size Planning

In general, given a planning value for σ or p (and ignoring the fpc), we can approximate the sample size necessary so that the half-width of the resulting interval will be H:

Parameter	Sample Size
μ	$n \geq \left[\frac{z(\alpha/2)\sigma}{H}\right]^2$
p	$n \geq \left[\frac{z^2(\alpha/2)p(1-p)}{H^2}\right]$

7.5 Stratified Random Sampling

In a stratified sample:

(1) The population is partitioned into L distinct strata

(2) Simple random samples are drawn from each stratum

Advantages of stratified sampling:

(1) Forces sampler to obtain information from every stratum; no such guarantees with s.r.s; and

(2) Better precision, so fewer observations needed

Stratified Random Sampling--Examples

(1) Auditing: stratify accounts based on book values

(2) In a study of shipments of household goods by motor carriers (truckers): stratify by carrier size--small, medium, large

(3) Political polling: stratify by geographic area, or socioeconomic class

Notation:

$$
\begin{aligned}
N_l &= \ l\text{th stratum population size} \\
n_l &= \ l\text{th stratum sample size} \\
N &= \ \text{population size} \\
&= \sum_{i=1}^{L} N_l \\
n &= \ \text{sample size} \\
&= \sum_{i=1}^{L} n_l \\
W_l &= \ \text{fraction of population in } l\text{th stratum} \\
&= \frac{N_l}{N}
\end{aligned}
$$

Parameters: Population Mean

$$
\begin{aligned}
\mu_l &= \ \text{population mean in the } l\text{th stratum} \\
\mu &= \ \text{population mean} \\
&= \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} x_{li} \\
&= \frac{1}{N} \sum_{l=1}^{L} N_l \mu_l \\
&= \sum_{l=1}^{L} W_l \mu_l
\end{aligned}
$$

Parameters: Population Variance

$$\sigma^2 = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(x_{li}-\mu)^2$$
$$= \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(x_{li}-\mu_l+\mu_l-\mu)^2$$
$$=$$

= ***within strata*** **variance +** ***between strata*** **variance**

Estimates

$$\bar{X}_l = \ l\text{th stratum sample mean}$$
$$= \frac{1}{n_l}\sum_{i=1}^{n_l} x_{il}$$
$$\bar{X}_S = \ \text{stratified-sampling-based estimate of } \mu$$
$$= \sum_{l=1}^{l} W_l \bar{X}_l$$
$$Var(\bar{X}_S) = \sum_{l=1}^{l} W_l^2 \sigma^2_{\bar{X}_l}$$
$$= \sum_{l=1}^{l} W_l^2 \left(\frac{\sigma_l^2}{n_l}\right)\left(1-\frac{n_l-1}{N_l-1}\right)$$

Estimated Variance

Substituting an unbiased estimate of s_l^2 into the above expression gives:

$$s_{\bar{X}}^2 = \sum_{l=1}^{l} W_l^2 \frac{1}{n_l}\left[s_l^2\left(\frac{N_l-1}{N_l}\right)\right]\left(1-\frac{n_l-1}{N_l-1}\right)$$
$$= \sum_{l=1}^{l} W_l^2 \frac{s_l^2}{n_l}\left(1-\frac{n_l}{N_l}\right)$$

For small sampling fractions,

$$s_{\bar{X}}^2 = \sum_{l=1}^{l} W_l^2 \frac{s_l^2}{n_l}$$

Example 7.5-1 Sampling of Hospitals

Objective is to estimate average number of discharges; we will assume that the number of beds in each hospital is known and that discharges is likely to be proportional to hospital size (number of beds).

Thus we will stratify hospitals by beds (here using equal-size strata):

Stratum	N_l	W_l	μ_l	σ_l
1	98	0.249	182.9	103.4
2	98	0.249	526.5	204.8
3	98	0.249	956.3	243.5
4	99	0.250	1591.2	419.2

Example 7.5-1 (Continued)

(a) Compute $Var(\bar{X}_S)$ and compare to $Var(\bar{X}_{srs})$

Example 7.5-1 (Continued)

(b) Estimate μ and $Var(\bar{X}_{srs})$ using the following sample results:

Stratum	X_l	s^2_l
1	240.6	6827.6
2	507.4	23790.7
3	865.1	42573.0
4	1716.5	152099.6

Example 7.5-1 (Continued)

(c) Give an approximate 95% confidence interval for μ

7.5.3 Methods of Allocation

1. Optimal allocation: Choose $n_1, \ldots, n_L$ to minimize $Var(\bar{X}_s)$, subject to the constraint that $\Sigma n_i = n$.

$$n_l = n\frac{W_l\sigma_l}{\sum_{k=1}^{L} W_k\sigma_k} \qquad \text{where } l = 1, \ldots, L$$

Also called Neyman allocation; found by introducing a Lagrange multiplier for the constraint and then setting partial derivatives of $Var(\bar{X}_s)$ to zero and solving.

Major disadvantage:

Variance of $\bar{X}$ with Optimal Allocation

Proportional Allocation

2. In *proportional allocation*, the strata sample sizes are taken in direct proportion to the strata population sizes:

$$n_l = n\frac{N_l}{N} = nW_l$$

Notes:

1) $\bar{X}_{sp}$ is just the unweighted sample average:

2) $Var(\bar{X}_{sp})$ =

Optimal vs Proportional vs s.r.s

The reduction in variance from use of optimal allocation over proportional allocation is:

$$Var(\bar{X}_{sp}) - Var(\bar{X}_{so}) = \frac{1}{n}\sum_{i=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$$

where:

$$\bar{\sigma} = \sum_{i=1}^{L} W_l \sigma_l$$

The reduction in variance from use of proportional allocation over s.r.s (no stratification) is:

$$Var(\bar{X}) - Var(\bar{X}_{sp}) = \frac{1}{n}\sum_{i=1}^{L} W_l(\mu_l - \mu)^2$$

Chapter 8: Parameter Estimation and Fitting Probability Distributions

Example 8.2-1 Fitting a Poisson Distribution

The data below show counts of alpha particles emitted from americum 241 in 1207 10-second intervals.

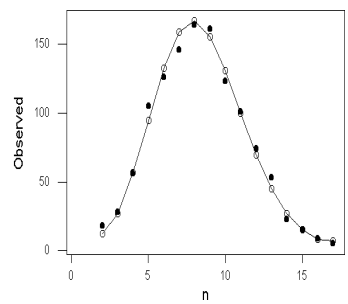
The average ($\bar{X}$) was 8.392

n	Observed
0 - 2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9
17 +	5
	1207

Using $\lambda = \bar{X}$...

We can compare the expected number of counts from Pois(λ) to what was observed, numerically, graphically:

n	Observed	Prob	Expected
0-2	18	0.010	12.20
3	28	0.022	26.95
4	56	0.047	56.54
5	105	0.079	94.90
6	126	0.110	132.73
7	146	0.132	159.12
8	164	0.138	166.92
9	161	0.129	155.65
10	123	0.108	130.62
11	101	0.083	99.65
12	74	0.058	69.69
13	53	0.037	44.99
14	23	0.022	26.97
15	15	0.012	15.09
16	9	0.007	7.91
17+	5	0.006	7.08
	1207	1.000	1207.00



Goodness of Fit

One measure of goodness-of-fit of the hypothesized Poisson(8.392) model is:

$$X^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(df)$$

if the Poisson model is correct, where:

df = # cells - # parameters estimated - 1

This is known as the Pearson Chi-square goodness-of-fit statistic.

Testing for Goodness of Fit

Step 1. Null hypothesis. H_0: $X \sim Pois(\lambda)$

Step 2. Alternative hypothesis. H_1: $X \nsim Pois(\lambda)$

Step 3. Test statistic is:

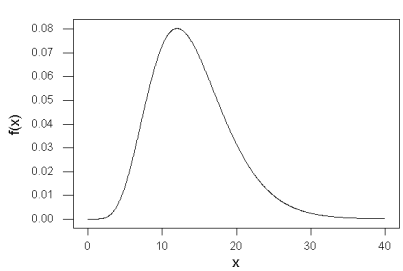
$$X^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

Step 4: Reject H_0 if X^2 is too large to have come from the distribution under the null hypothesis

What do you mean, too large?

Use p-value:

$$\text{p-value} = P(\chi^2 > X^2)$$



Reject H_0 if X^2 is large, or p-value is small, say < .05 or .01

Testing for Goodness of Fit

Step 1. Null hypothesis. H_0: X ~ Pois(8.392)

Step 2. Alternative hypothesis. H_1: X $\not\sim$ Pois(8.392)

Step 3. Test statistic is:

$$X^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i} = 8.99$$

Step 4: p-value = $p(\chi^2(14) > 8.99) = .83$; Nothing odd about 8.99, fail to reject H_0

8.4 Method of Moments

The Poisson example used the "method of moments" to fit a Poisson distribution to the data.

Definition. The kth sample moment is:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} X_i^k}{n}$$

Since $\hat{\mu}_k$ converges in probability to $\mu_k = E(X^k)$, we can use $\hat{\mu}_k$ as an estimate of μ_k.

Method of Moments for Two Parameters

Suppose we wish to estimate parameters θ_1 and θ_2, where:

$$\theta_1 = g_1(\mu_1, \mu_2)$$
$$\theta_2 = g_2(\mu_1, \mu_2)$$

Then the method of moments estimators (MMEs) of θ_1 and θ_2 are:

$$\hat{\theta}_1 = g_1(\hat{\mu}_1, \hat{\mu}_2)$$
$$\hat{\theta}_2 = g_2(\hat{\mu}_1, \hat{\mu}_2)$$

Example 8.4-1 Poisson Distribution
We wish to estimate $\theta_1 = \lambda$.:

MME Examples

Example 8.4-2 Normal Distribution
We wish to estimate μ and σ^2

MME Examples

Example 8.4-3 X ~ Gamma(α, λ). Give MME estimators for α and λ.

$$\mu_1 = \frac{\alpha}{\lambda} = g_1^{-1}(\alpha, \lambda)$$

$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2} = g_2^{-1}(\alpha, \lambda)$$

Solving for α and λ, we have:

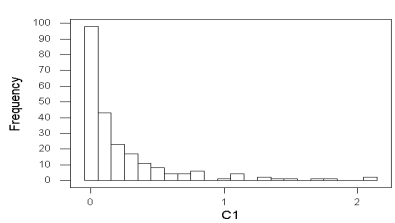
$$\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2} = g_1(\mu_1, \mu_2)$$

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} = g_2(\mu_1, \mu_2)$$

so the MMEs are:

Rainfall Data

The histogram below summarizes the amount of rain in each of 227 storms during 1960-1964 in a metropolitan area



To fit a gamma density to these data, the first and second sample moments are .224 and .184, respectively. Therefore, $\hat{\alpha}$ = .375, $\hat{\lambda}$ = 1.674

Method of Moments: General Case

Step 1: **Calculate low-order moments, finding expressions for the moments in terms of the parameters**

Step 2: **Invert the expressions found in Step 1, finding new expressions for the parameters in terms of the moments**

Step 3: **Insert the sample moments into the expressions obtained in Step 2, thus obtaining estimates of the parameters**

MMEs are "Consistent" Estimators

Definition: Let $\hat{\theta}_n$ be an estimate of θ based on a sample of size n. $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ: for any $\varepsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

Result: If the functions g_i are continuous, the MMEs are consistent.

Parametric Bootstrap Estimates of Standard Errors

When no functional form exists for the standard deviation of an estimate, we can use simulation.

Step 1: **Generate 1000 (say) samples of size n from the estimated distribution, and compute the MME for each sample.**

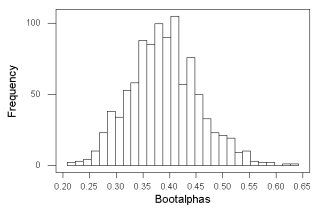
Step 2: **Compute the sample standard deviation of the 1000 MMEs.**

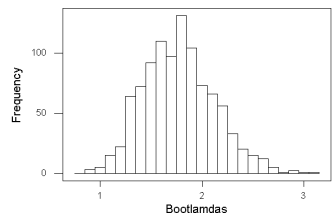
A plot of the 1000 values gives the "bootstrap distribution"

Minitab Exec for Parametric Bootstrap

```
let k1 = k1 + 1
random 227 c2;
  gamma .375 .597.
let k2 = mean(c2)
let k3 = mean(c2**2)
let k4 = k2**2/(k3-k2**2)
let k5 = k2/(k3-k2**2)
let c3(k1) = k4
let c4(k1) = k5
```

For the Rainfall Data






```
MTB > let k2 = stdev(c3)
MTB > let k3 = stdev(c4)
MTB > print k2 k3
```

Data Display

```
K2    0.0647840
K3    0.347379
```


8.5 Method of Maximum Likelihood

General, flexible, widely-used method for estimating parameters. MLEs have excellent statistical properties.

1. Likelihood

Suppose $X_1, \ldots, X_n$ have a joint density or frequency function $f(x_1, \ldots, x_n | \theta)$. After we have collected the data, the observations $x_1, \ldots, x_n$ are known; only the parameter θ is unknown. Plugging in the known values we obtain the likelihood function:

$$\text{lik}(\theta) = f(x_1, \ldots, x_n | \theta).$$

ML Estimation (Continued)

If the X_is are *iid* (the usual case), then:

$$lik(\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

2. Log-likelihood

Working with the log of the likelihood (which is a product) is usually easier to work with:

$$l(\theta) = \ln[\text{lik}(\theta)] = \sum_{i=1}^{n} \ln[f_X(x_i|\theta)]$$

Maximum Likelihood Estimation

Definition: The *maximum likelihood estimate* (MLE) of θ is the value of θ that maximizes the likelihood function (or equivalently, the log-likelihood function).

Intuition: The MLE is the value of the parameter(s) that maximize(s) the probability of the data that we observed

Examples

Example 8.5.-1 Suppose Xi ~ Bernoulli(p).

a) Give the likelihood for n observations with k ones and (n - k) zeros

b) Give the log likelihood

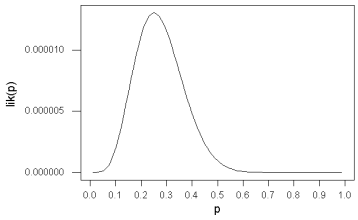
c) Plot the likelihood and log likelihood for n = 20, k = 5

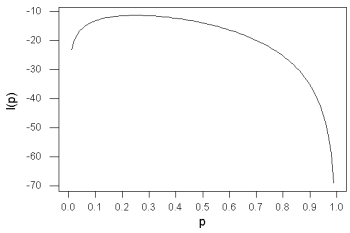
d) Find the MLE using calculus

Examples

Example 8.5.-1 (Continued)

c) Plot the likelihood and the log-likelihood





Examples

Example 8.5.-1 (Continued)

d) Use calculus to find the MLE of p.

Examples

Example 8.5.-2 Suppose $X_1, \ldots, X_n$ are *iid* Poisson(λ).

a) Give the likelihood

b) Give the log likelihood

c) Plot the log likelihood for n = 5, and $x_1 = 2$, $x_2 = 8$, $x_3 = 5$, $x_4 = 3$, and $x_5 = 2$.

d) Find the MLE using calculus

---

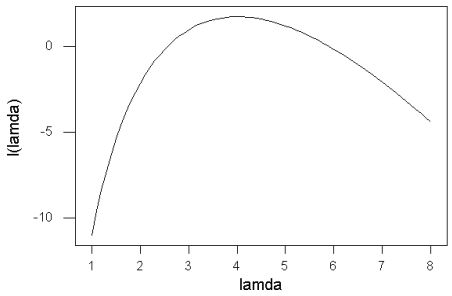
a)

b)

Examples

Example 8.5-2 (Continued)

c) Plot the log-likelihood for the data given



Examples

Example 8.5-2 (Continued)

c) Find the expression for the MLE

Examples

Example 8.5-3 Normal Distribution. Suppose $X_1, \ldots, X_n$ are *iid* Normal(μ, σ^2).

a) Give the likelihood

b) Give the log likelihood

c) Find the MLE using calculus

Examples

Example 8.5-3 Normal Distribution (continued)

8.5.1 ML Estimation of Multinomial Cell Probabilities

Example 8.5-4 (Mendel's Peas)

In a famous experiment, Gregor Mendel crossed 556 smooth, yellow, male peas with wrinkle, green, female peas. The observed counts were as follows:

Cell	Type	Observed Count
1	Smooth yellow	315
2	Smooth green	108
3	Wrinkled yellow	102
4	Wrinkled green	31
		556

Estimate the cell probabilities using maximum likelihood.

Mendel's Peas (Continued)

8.5.2 Large Sample Theory for Maximum Likelihood Estimation

Basic Results:

1) **MLEs are consistent**
2) **MLEs are asymptotically normal**
3) **MLEs are asymptotically unbiased**
4) **MLEs are asymptotically efficient**

Result 1. Under appropriate smoothness conditions on the density f, the MLE from an iid sample is consistent.

MLE Consistency Proof

Let θ_0 denote the true parameter. Consider maximizing:

$$\begin{aligned}\frac{1}{n}l(\theta) &= \frac{1}{n}\sum_{i=1}^{n} log f(x_i|\theta) \\ &\rightarrow E[log f(X|\theta)] \\ &= \int log[f(X|\theta)]f(X|\theta_0)dx\end{aligned}$$

To maximize the above, take the derivative, interchange the order of differentiation and integration, and set to zero:

$$\frac{\partial}{\partial\theta}\int log[f(x|\theta)]f(x|\theta_0)dx = \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}f(x|\theta_0)dx = 0$$

What value of θ will solve this equation? As a guess at the maximizing value, try θ_0:

MLE Consistency Proof (continued)

Then:

$$\int \frac{\frac{\partial}{\partial\theta} f(x|\theta_0)}{f(x|\theta_0)} f(x|\theta_0)dx = \int \frac{\partial}{\partial\theta} f(x|\theta_0)dx$$
$$= \frac{\partial}{\partial\theta}\int f(x|\theta_0)dx$$
$$= \frac{\partial}{\partial\theta}(1) = 0 \qquad \text{True!}$$

Thus the MLE $\hat{\theta}$ converges in probability to θ_0. □

Information Function

Definition. The "information" of θ is defined as:

$$I(\theta) = E[\frac{\partial}{\partial\theta} log\, f(X|\theta)]^2$$

Lemma.

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} log f(X|\theta)]$$

Justification: Since ∫ f(x|θ)dx = 1,

$$\frac{\partial}{\partial\theta}\int f(x|\theta)dx = 0$$

Also, since

Information Function (continued)

$$\frac{\partial}{\partial\theta} f(x|\theta) = \left[\frac{\partial}{\partial\theta} log f(x|\theta)\right] f(x|\theta)$$

we have:

$$0 = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \int \left[\frac{\partial}{\partial\theta} log f(x|\theta)\right] f(x|\theta)dx$$

Taking derivatives again,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta} log f(x|\theta)\right] f(x|\theta)dx \\ &= \int \frac{\partial^2}{\partial\theta^2} log[f(x|\theta)]f(x|\theta) + \int \left[\frac{\partial}{\partial\theta} log f(x|\theta)\right] \frac{\partial}{\partial\theta} f(x|\theta)dx \\ &= \int \frac{\partial^2}{\partial\theta^2} log[f(x|\theta)]f(x|\theta) + \int \left[\frac{\partial}{\partial\theta} log f(x|\theta)\right]^2 f(x|\theta)dx \end{aligned}$$

Information Function (continued)

which implies that:

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} log f(X|\theta)] \qquad \square$$

Result (2). The large sample distribution of an MLE is approximately normal with mean θ_0 (the true parameter) and variance $1/[nI(\theta_0)]$.

Justification: From a Taylor series expansion,

$$0 = l'(\hat{\theta}) \approx l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_0)$$

Proof of Asymptotic Normality

Solving for the MLE, we have:

$$\hat{\theta} \approx \frac{-l'(\theta_0)}{l''(\theta_0)} + \theta_0$$

Now:

$$\begin{aligned}
\frac{1}{n} l''(\theta_0) &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial^2}{\partial \theta^2} log f(x|\theta) \right]_{\theta=\theta_0} \\
&\to E\left[\frac{\partial^2}{\partial \theta^2} log f(x|\theta) \right] \\
&= -I(\theta_0) \qquad \text{by the LLNs, so} \\
l''(\theta_0) &\to -nI(\theta_0)
\end{aligned}$$

Therefore: $$\hat{\theta} \approx \frac{-l'(\theta_0)}{-nI(\theta_0)} + \theta_0$$

Proof of Asymptotic Normality (cont)

1. **Expectation:**

$$\begin{aligned}
E(\hat{\theta}) &= \frac{1}{nI(\theta_0)} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta} log f(x_i|\theta) \right]_{\theta=\theta_0} + \theta_0 \\
&= \theta_0 \qquad \text{(see consistency proof)}
\end{aligned}$$

2. **Variance:**

$$\begin{aligned}
Var(\hat{\theta}) &= \left[\frac{1}{nI(\theta_0)} \right]^2 Var\left(\sum_{i=1}^{n} \left[\frac{\partial}{\partial \theta} log f(x_i|\theta) \right]_{\theta=\theta_0} \right) \\
&= \frac{1}{n^2 I^2(\theta_0)} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta} log f(x_i|\theta_0) \right]^2 \\
&= \frac{1}{n^2 I^2(\theta_0)} nI(\theta_0) \\
&= \frac{1}{nI(\theta_0)}
\end{aligned}$$

Proof of Asymptotic Normality (cont)

3. Normality

$$\hat{\theta} = \frac{1}{nI(\theta_0)} \sum_{i=1}^{n} \left[\frac{\partial}{\partial \theta} log f(x_i|\theta) \right]_{\theta=\theta_0} + \theta_0$$

$$= \frac{1}{nI(\theta_0)} \sum_{i=1}^{n} D_i$$

where $D_1, \ldots, D_n$ are iid. Therefore, for large n:

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

8.5.3 Confidence Intervals for MLEs

1. Approximate confidence intervals using large sample theory (asymptotic normality).

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{1}{nI(\hat{\theta})}}$$

Example 8.5-5 Assume that $X_1, \ldots, X_n$ are Pois(λ). Give a formula for a (1-α)100% confidence interval for λ.

Example 8.5-5 (continued)

2. Parametric Bootstrap Intervals

Step 1. Estimate θ with $\hat{\theta}$.

Step 2. Generate 1000 samples of size n from $f(x|\hat{\theta})$ (equivalently, from $F(x|\theta)$). Compute the MLE θ_i^* for each sample i.

Step 3. Compute the sample standard deviation of the 1000 MLEs:

$$s_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{1000}(\theta_i^* - \bar{\theta}^*)^2}{1000-1}}$$

Step 4. Compute the lower and upper 90% limits by the *reflection method*:

$$LL = \hat{\theta} - (\theta^*_{(950)} - \hat{\theta}) \qquad UL = \hat{\theta} - (\theta^*_{(50)} - \hat{\theta})$$

Example 8.5-6. Parametric Bootstrap

Consider the following sample of size n = 16 from a Poisson distribution: 4, 6, 4, 11, 3, 5, 6, 7, 5, 8, 3, 6, 8, 5, 6, 7, for which $\bar{X}$ = 5.875. Find a 90% bootstrap interval estimate for λ.

Calling Exec: Poisson Bootstrap 1.txt

```
let k1 = 0
let k2 = mean(c1)
exec 'c:\files\teaching\6850--Math Stats\Poisson Bootstrap 2.txt' 1000
sort c3 c3;
 by c3.
name k3 'LL'
name k4 'UL'
let k3 = mean(c1) - (c3(950) - mean(c1))
let k4 = mean(c1) - (c3(50) - mean(c1))
print k3 k4
```


Example 8.5-6. Parametric Bootstrap

Called Exec: Poisson Bootstrap 2.txt

```
let k1 = k1 + 1
random 16 c2;
  Poisson k2.
let c3(k1) = mean(c2)
```

Results:

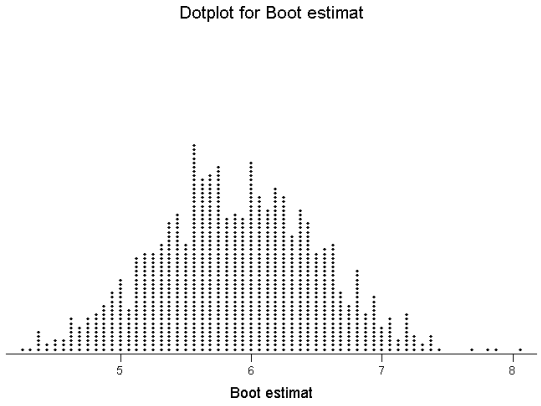
```
MTB > Execute "C:\FILES\Teaching\6850--Math Stats\Poisson Bootstrap 1.txt" 1.
Executing from file: C:\FILES\Teaching\6850--Math Stats\Poisson Bootstrap 1.txt
Executing from file: c:\files\teaching\6850--Math Stats\Poisson Bootstrap 2.txt
```

Data Display

```
LL     4.81250
UL     6.87500
```

Results

Dot plot of bootstrap MLEs:



Non-parametric Bootstrap

In the parametric bootstrap, we assumed that we knew the CDF F (Poisson). What if we haven't a clue?

For example, suppose we would like to develop a 90% confidence interval for the true median (50th percentile) but we haven't any idea about the true distribution of the data.

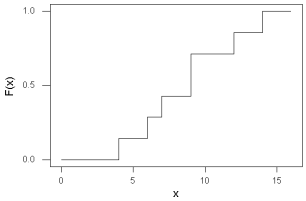
How can we do a bootstrap in this situation?

Empirical CDF

Answer: Estimate F using F_n, the *empirical CDF*, and then sample from F_n. F_n is the discrete distribution that places probability 1/n on each of the n observed values.

Example 8.5-7: Suppose the sample consisted of the 7 values: 4, 6, 7, 9, 9, 12, 14. Then the empirical CDF is determined as follows:

x_i	$P(x_i)$	$F(x_i)$
4	1/7	1/7
6	1/7	2/7
7	1/7	3/7
9	2/7	5/7
12	1/7	6/7
14	1/7	7/7



How Do You Sample From F_n?

Sampling from F_n is equivalent to sampling *with replacement* from the set of sample items $\{x_1,\ldots,x_n\}$!

Result: Non-parametric Bootstrap.

Procedure is exactly the same as on page 368 for the parametric bootstrap, except that we sample from Fn rather than the known distribution F. The "called" exec becomes:

```
let k1 = k1 + 1
sample 7 c1 c2;
  replace.
let c3(k1) = median(c2)
```

Non-parametric Bootstrap Example

Example 8.5-8. For the data in Example 8.5-8, generate a 95% bootstrap confidence interval for the true median.

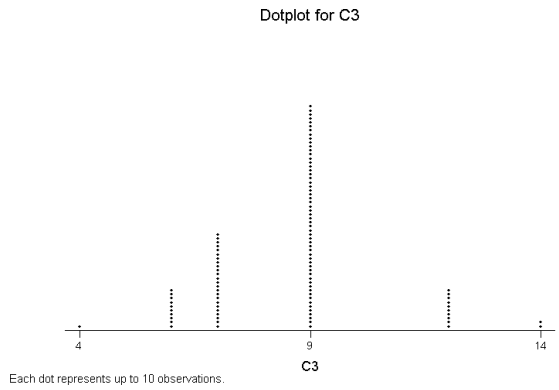
```
MTB > Execute "C:\FILES\Teaching\6850--Math Stats\Median Bootstrap 1.txt" 1.
Executing from file: C:\FILES\Teaching\6850--Math Stats\Median Bootstrap 1.txt
Executing from file: c:\files\teaching\6850--Math Stats\Median Bootstrap 2.txt
```

Data Display

```
LL    6.00000
UL    12.0000
```


Non-parametric Bootstrap Example

Plot of bootstrap medians:



8.6 Efficiency and the Cramer-Rao Lower Bound

Suppose you have two estimators, for example, $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MME}$. Which is better? How to compare?

1. Use the mean squared error:

$$\frac{MSE(\hat{\theta}_{MLE})}{MSE(\hat{\theta}_{MME})} = \frac{Var(\hat{\theta}_{MLE}) + [Bias(\hat{\theta}_{MLE})]^2}{Var(\hat{\theta}_{MME}) + [Bias(\hat{\theta}_{MME})]^2}$$

2. If both are unbiased, then:

$$\frac{MSE(\hat{\theta}_{MLE})}{MSE(\hat{\theta}_{MME})} = efficiency(\hat{\theta}_{MLE}, \hat{\theta}_{MME})$$
$$= \frac{Var(\hat{\theta}_{MLE})}{Var(\hat{\theta}_{MME})}$$

Cramer-Rao Inequality

For any unbiased estimator $\hat{\theta}_{UB}$ of θ,

$$MSE(\hat{\theta}_{UB}) = Var(\hat{\theta}_{UB}) \geq \frac{1}{nI(\theta)}$$

Proof: See text, p. 275.

Definition: An unbiased estimator is *efficient* if

$$Var(\hat{\theta}_{UB}) = \frac{1}{nI(\theta)}$$

Result: MLEs are asymptotically efficient! (See p 364).

MLE Wrap-up

Notes:

1. **For finite n, it may be that the MLE is neither unbiased, nor efficient; results reported are asymptotic (large sample) results!!!!!**

2. **Simulation (bootstrapping) can help determine actual bias, variance**

3. **Summary: MLEs are asymptotically unbiased, efficient, and have a normal sampling distribution with variance $[nI(\theta)]^{-1}$**

Chapter 9: Testing Hypotheses and Assessing Goodness of Fit

9.1 Introduction

Example 9.1-1

Suppose a friend claims to have extra-sensory perception (ESP) and you wish to test this claim. To do so, you will show your friend 10 cards (backside only, not revealing the face) from a well-shuffled deck and ask whether the card is black or red.

If your friend is guessing, the probability of a correct guess is .5. If your friend has ESP, the probability will be larger than .5 (not necessarily 1.0).

Null and Alternative Hypotheses

Hypothesis: A statement about a probability distribution

- **"Simple" if it completely specifies the distribution**
  Example: X ~ Binomial(.5, 10)
- **"Composite" if it does not completely specify distribution**
  Example: X ~ Binomial(p, 10), where p > .5

Null hypothesis: H_0

- **A claim that cannot be proven, only disproved**
- **Status quo; generally accepted, assumed to be true**

Alternative hypothesis: H_A (sometimes written H_1)

- **Hypothesis to be proven**
- **Often the opposite of H_0, but not always**

Four-Step Hypothesis Testing Framework

In the ESP example:

1. Null hypothesis: **H_0:**

2. Alternative hypothesis: **H_A:**

3. Test Statistic: **T, total correct out of 10**

4. Decision rule: **Reject H_0 if T > 7**

Note: ***Rejection Region*** **is {8, 9, 10}**
Acceptance Region **is {0, 1, 2, 3, 4, 5, 6, 7}**

Errors in Hypothesis Testing

	State of Nature	
Action	**H_0 True**	**H_0 False**
Accept H_0	No Error	Type II Error
Reject H_0	Type I Error	No Error

- **P(Type I Error) = P(reject H_0|H_0 true) = α**
- **P(Type II Error) = P(accept H_0|H_0 false) = β**
- **Power of the test = P(reject H_0|H_0 false) = 1 - β**

Finding a Type I Error Rate

Example 9.2-1 Find the probability of a Type I error in ESP Example 9.1-1.

α = P(Type I Error) = P(reject H_0|H_0 true)

Finding a Type II Error Rate and Power

To compute a Type II error, both hypotheses must be simple.

Example 9.2-2 Find the probability of a Type II error in ESP Example 9.1-1, under the alternative that p = .6. What is the power of the test for this alternative?

β = P(Type II Error) = P(accept $H_0 | H_A$ true)

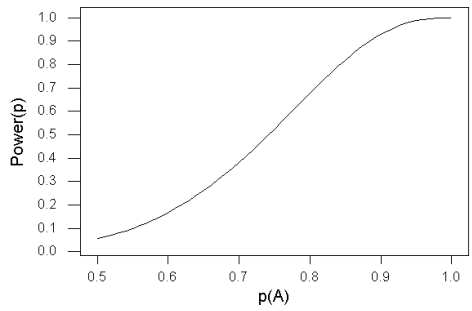
Finding a Type II Error Rate and Power

The Type II error rate will change as the alternative hypothesis changes. Smaller differences between H_0 and H_A are harder to detect.

Example 9.2-3 Find the power and the probability of a Type II error in ESP Example 9.1-1, under the alternative p = .55.

Power (Characteristic) Curve for the Test

Plot the power of the test vs the parameter p in the alternative (A) distribution



p-value of a Test

Suppose that:

- **large values of the test statistic, T, support H_A**
- **small values support H_0**
- **the specific value of the test statistic we observed was T***

Then

1. **p-value = $P(T > T^* | H_0 \text{ is true})$**
2. **H_0 is rejected if p-value < α**

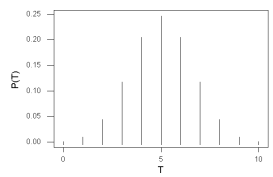
Note: P is sometimes defined to be the probability of obtaining a test statistic that is as contradictory or more contradictory to H_0 as the observed test statistic)

Finding the p-value of a Test

Suppose that in the ESP test the observed number correct turned out to be T* = 9. Then

$$\begin{aligned} \text{p-value} &= P(T \geq 9 \mid H_0 \text{ is true}) \\ &= P(T \geq 9 \mid p = 5) \\ &= .0107 \end{aligned}$$

So, if H_0 is true, a value as large or larger than 9 will happen only 1% of the time. Picture:



9.3 Likelihood Ratio Tests for Simple Hypotheses: The Neyman-Pearson Lemma

We can measure the relative plausibility (likelihood) of the null and alternative hypotheses by computing the ratio of the likelihoods of the observed data X under the null and alternative distributions:

$$LR = \frac{f_0(X)}{f_A(X)}$$

Here, small values support H_A, large ones support H_0.

Example 9.3-1. Calculate the likelihood ratio (LR) for the simple null hypothesis that p = .5 and the simple alternative that p = .6, if X = 4.

$$LR = \frac{f_0(4)}{f_A(4)} = \frac{\binom{10}{4}.5^4.5^6}{\binom{10}{4}.6^4.4^6} = \frac{.2051}{.1115} = 1.84$$

The Neyman-Pearson Lemma

Suppose the LR test (which rejects H_0 for small values of LR) has significance level α. Then any other test which has significance level $\alpha^* \leq \alpha$ has power less than or equal to that of the LR test.

Example 9.3-2. Develop a LR test for the ESP example.

LR	X
9.31	0
6.21	1
4.14	2
2.76	3
1.84	4
1.23	5
0.82	6
0.55	7
0.36	8
0.24	9
0.16	10

Since small values of LR correspond to large values of X, we can make the rejection region correspond to large values of X. α can be determined since we know the distribution of T = X

Example 9.3-3 Testing a Normal Mean

In the production of Cheerios (breakfast cereal), the production process is set to fill each 18 ounce box with 18.2 ounces, on average. This is to ensure that the customer receives at least 18 ounces as claimed on the box. Suppose it is known from previous data that the standard deviation of box weights is .06 ounces, and that box weights are approximately normally distributed.

Null hypothesis:

Alternative hypothesis:

Develop an LR test for these alternatives

Example 9.3-3 Testing a Normal Mean

$$LR = \frac{f_0(X)}{f_A(X)} = \frac{\frac{1}{\sqrt{2\pi}(.06)} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 18.2}{.06}\right)^2}}{\frac{1}{\sqrt{2\pi}(.06)} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - 18.0}{.06}\right)^2}}$$

Example 9.3-3 (continued)

Can use Z to determine a rejection region for α based on small values of Z.

Example 9.3-4 Cheerios Again

Suppose, in Example 9.3-4, that n = 9 and $\bar{X}$ = 18.15. Test the hypothesis that μ = 18.2 versus the alternative that μ = 18.0 at the α = .05 level of significance.

9.5 Generalized Likelihood Ratio Tests

What is the main limitation with LR tests?

Generalized likelihood ratio **tests are to testing, as maximum likelihood estimation is to estimation:**

- **You can almost always construct these**
- **They go hand-in-hand with maximum likelihood**
- **Composite hypotheses are not a problem.**

The null and alternative hypotheses are written is a slightly different form, using set notation.

Set Notation for GLR Tests

Suppose one or both of the hypotheses are composite:

Standard	Set Notation
H_0: $\mu = \mu_0$	H_0: $\mu \in \omega_0 = \{\mu_0\}$
H_A: $\mu \neq \mu_0$	H_A: $\mu \in \omega_A = \{\mu \mid \mu \neq \mu_0\}$

In general, for GLR tests, we will use set notation in a slightly different form than above:

H_0: $\mu \in \omega_0 = \{\mu_0\}$

H_A: $\mu \in \omega_0 \cup \omega_A = \Omega = \{\mu \mid -\infty < \mu < \infty\}$

Set Notation for GLR Tests

So, the null hypothesis corresponds to the parameter μ (or parameter vector θ) being restricted to a reduced space ω_0, while the alternative hypothesis indicates that the parameter (vector) is an element of a larger space, Ω.

Notes:

1. **Dimension of ω_0 = number of free parameters in ω_0**
2. **Dimension of Ω = number of free parameters in Ω**
3. **Test degrees of freedom DF = Dim Ω – Dim ω_0**
3. **Usually, $\omega_0 \subset \Omega$.**

Generalized Likelihood Ratio Test

1. **Null hypothesis:** $\theta \in \omega_0$

2. **Alternative hypothesis:** $\theta \in \Omega$

3. **Test statistic:**

$$GLR = \Lambda = \frac{max_{\ \theta \in \omega_0}[lik(\theta)]}{max_{\ \theta \in \Omega}[lik(\theta)]}$$

$$T^* = -2log\,\Lambda \sim \chi^2_{DF}$$

4. **Decision rule: p-value = P{T > T*} < α, reject H_0**

Generalized Likelihood Ratio Null Distribution

Under smoothness conditions on the probability density or frequency functions involved, the null distribution of -2 log Λ tends to a chi-square distribution with degrees of freedom equal to Dim Ω – Dim ω_0.

Summary/Implications

Any time you estimate parameters via maximum likelihood, a large sample (approximate) test is available:

Step 1: Find the value of $\hat{\theta}_0$ that maximizes the likelihood for $\theta \in \omega_0$

Step 2: Find the value of $\hat{\theta}_A$ that maximizes the likelihood for $\theta \in \Omega$

Step 3: Compute: $$\Lambda = \frac{[lik(\hat{\theta}_0)]}{[lik(\hat{\theta}_A)]}$$

Step 4: Compute: $T^* = -2 \log \Lambda \sim \chi^2(DF)$

Example 9.5-1 (Mendel's Peas, Again)

According to genetic theory, the relative frequencies of the four types of peas should be as listed below. Multiplying by n = 556 gives the expected frequencies.

Cell	Type	Observed Count	Relative Frequencies	Expected Frequencies
1	Smooth yellow	315	9/16	312.75
2	Smooth green	108	3/16	104.25
3	Wrinkled yellow	102	3/16	104.25
4	Wrinkled green	31	1/16	34.75
		556		

Find the GLR test.

Example 9.5-1 (Mendel's Peas, Again)

H_0: $p \in \{p_1 = 9/16, p_2 = 3/16, p_3 = 3/16, p_4 = 1/16\}$

H_A: $p \in$ three-dimensional simplex defined by:
$0 \le p_1 \le 1, 0 \le p_2 \le 1, 0 \le p_3 \le 1, 0 \le p_4 \le 1,$
and $\Sigma\, p_i = 1.$

Under H_0:

Under H_A:

Example 9.5-1 (Mendel's Peas, continued)

Example 9.5-2 (Mendel's Peas, Again)

Test Mendel's data again for goodness of fit to the hypothesized distribution, this time using the Pearson chi-square test (see slide page 331).

$$X^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

Where DF = # cells - # parameters estimated - 1
= 4 - 0 - 1 = 3.

Here X^2 = .604, and p-value = $P(\chi^2_3 > .604)$ = .8955

Another Multinomial Test: Chi-square Test of Independence

Example 9.5-3 An examination of a sample of 183 tenure cases during the past five years at a major university yielded the following 2-by-2 contingency table:

	Tenure		
Gender	Granted	Rejected	Total
Males	115	22	137
Females	31	15	46
Total	146	37	183

Is the granting of tenure independent of gender? (Are males more likely to receive tenure than females?)

General I by J Contingency Table Setup

An examination of a sample of 183 tenure cases during the past five years at a major university yielded the following 2-by-2 contingency table:

	Factor B Categories				
Factor A Categories	**1**	**2**	**…**	**J**	**Total**
1	n_{11}	n_{12}	…	n_{1J}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2J}	$n_{2\cdot}$
…	…	…		…	
I	n_{I1}	n_{I2}	…	n_{IJ}	$n_{I\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot J}$	$n_{\cdot\cdot}$

If A and B are independent, the joint probability p_{ij} of the ijth cell is the product of the marginal probabilities:

Null and Alternative Hypotheses

H_0: $p_{ij} = p_i \times p_j$ (A and B are independent)

$$p_{ij} \in \omega_0 = \{p_{ij} | p_{ij} = p_i p_j, \sum_{i=1}^{I} p_i = 1, \sum_{j=1}^{J} p_j = 1\}$$
$$\text{Dim } \omega_0 = (I-1) + (J-1)$$
$$\text{MLEs are: } \hat{p}_{ij} = \hat{p}_i \hat{p}_j = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \times \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

H_A: p_{ij} are free subject to $\Sigma\Sigma p_{ij} = 1$ (A, B are dependent)

$$p_{ij} \in \Omega = \{p_{ij} | \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1\}$$
$$\text{Dim } \Omega = IJ - 1$$
$$\text{MLEs are: } \hat{p}_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}$$

Test Statistic

Under H_0, the generalized likelihood ratio test statistic: follows a chi-square distribution:

$$X^2 = -2 log \Lambda$$
$$= -2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} log \frac{O_{ij}}{E_{ij}}$$

follows a chi-square distribution with:

$$O_{ij} = n_{ij} \text{ and } E_{ij} = \hat{p}_{ij} n_{..} = \frac{n_{i.} n_{.j}}{n_{..}}$$

and

$$df = \text{Dim}\,\Omega - \text{Dim}\,\omega_0$$
$$= (IJ - 1) - [(I - 1) + (J - 1)]$$
$$= (I - 1)(J - 1)$$

Alternatively, we can use the Pearson chi-square statistic.

Returning to the Example:

The fitted probabilities, observed and expected cell counts, and test statistics are:

i	j	Oij	pij	pi	pj	pipj	Eij	GLR	Pearson
1	1	115	0.628	0.749	0.798	0.597	109.301	11.691	0.297
1	2	22	0.12	0.749	0.202	0.151	27.699	-10.136	1.173
2	1	31	0.169	0.251	0.798	0.201	36.699	-10.464	0.885
2	2	15	0.082	0.251	0.202	0.051	9.301	14.339	3.493
								5.43	5.848

The critical region is: $X^2 > \chi^2_1(.95) = 3.84$, and the p-values are .0198 and .0156 for the GLR and Pearson chi-square tests.

Caveats: Pearson's works better for contingency tables.

Also, you need $E_{ij} > 5$ to get good results.

11.2.1 Some Common Tests Based on the Normal Distribution

1) One-sample t test of an hypothesized mean
2) Two-sample t test for differences in means
   - Equal variances
   - Unequal variances
3) One-sample chi-square test of an hypothesized variance
4) Two-sample F test for differences in variances

When normality assumption is not met, two options:

1) Transform the data (Box-Cox transformation)
2) Use a bootstrap-based test

1. One-Sample t test (Z test if σ is known)

Assumption: $X_1,\ldots,X_n$ are iid $N(\mu, \sigma^2)$.

1) H_0: $\mu = \mu_0$
2) H_A: (a) $\mu \neq \mu_0$, (b) $\mu > \mu_0$, or (c) $\mu < \mu_0$
3) Test statistic:

$$t^* = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

4) Critical regions:

$$\text{(a) } |t^*| > t(1 - \frac{\alpha}{2}; n-1)$$
$$\text{(b) } t^* > t(1-\alpha; n-1)$$
$$\text{(c) } t^* < -t(1-\alpha; n-1)$$

2. Two-Sample t test with Equal Variances

Assumptions: $X_{1,1},\ldots,X_{1,n1}$ are iid $N(\mu_1, \sigma^2)$.

$X_{2,1},\ldots,X_{2,n2}$ are iid $N(\mu_2, \sigma^2)$.

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

---

1) H_0: $\mu_1 = \mu_2$, or equivalently, $\mu_1 - \mu_2 = 0$
2) H_A: (a) $\mu_1 - \mu_2 \neq 0$, (b) $\mu_1 - \mu_2 > 0$, or (c) $\mu_1 - \mu_2 < 0$
3) Test statistic:

$$t^* = \frac{\bar{X}_1 - \bar{X}_2 - 0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

4) Critical regions:

(a) $|t^*| > t(1-\frac{\alpha}{2}; n_1+n_2-2)$

(b) $t^* > t(1-\alpha; n_1+n_2-2)$

(c) $t^* < -t(1-\alpha; n_1+n_2-2)$

Proof that $t^* \sim t(n_1 + n_2 - 2)$

From above,

$$t^* = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Approximate Two-Sample t test--Unequal Variances

Assumptions: $X_{1,1},\ldots,X_{1,n1}$ are iid $N(\mu_1, \sigma_1^2)$.
$X_{2,1},\ldots,X_{2,n2}$ are iid $N(\mu_2, \sigma_2^2)$.

$$df = \left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2 \div \left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}\right] \text{ (rounded to nearest integer)}$$

1) H_0: $\mu_1 = \mu_2$, or equivalently, $\mu_1 - \mu_2 = 0$
2) H_A: (a) $\mu_1 - \mu_2 \neq 0$, (b) $\mu_1 - \mu_2 > 0$, or (c) $\mu_1 - \mu_2 < 0$
3) Test statistic:

$$t^* = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(df)$$

4) Critical regions:
   (a) $|t^*| > t(1 - \frac{\alpha}{2}; df)$
   (b) $t^* > t(1 - \alpha; df)$
   (c) $t^* < -t(1 - \alpha; df)$

3. One-Sample Chi-square Test of Hypothesized Variance

Assumption: $X_1,\ldots,X_n$ are iid $N(\mu, \sigma^2)$.

1) H_0: $\sigma^2 = \sigma_0^2$
2) H_A: (a) $\sigma^2 \neq \sigma_0^2$, (b) $\sigma^2 > \sigma_0^2$, or (c) $\sigma^2 < \sigma_0^2$
3) Test statistic:

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

4) Critical regions:
   (a) $X^2 < \chi^2(\alpha/2; n-1)$ or $X^2 > \chi^2(1 - \alpha/2; n-1)$
   (b) $X^2 > \chi^2(1 - \alpha; n-1)$
   (c) $X^2 < \chi^2(\alpha; n-1)$

4. Two-Sample F Test for Differences in Variances

Assumptions: $X_{1,1},\ldots,X_{1,n1}$ are iid $N(\mu_1, \sigma^2)$.
$X_{2,1},\ldots,X_{2,n2}$ are iid $N(\mu_2, \sigma^2)$.

1) H_0: $\sigma_1^2 = \sigma_2^2$, or equivalently, $\sigma_1^2/\sigma_2^2 = 1$
2) H_A: (a) $\sigma_1^2/\sigma_2^2 \neq 1$, (b) $\sigma_1^2/\sigma_2^2 > 1$, or (c) $\sigma_1^2/\sigma_2^2 < 1$
3) Test statistic:

$$F^* = \frac{s_1^2}{s_2^2}$$

4) Critical regions:
   (a) $F^* < F(\alpha/2; n_1 - 1, n_2 - 1)$ or $F^* > F(1-\alpha/2; n_1 - 1, n_2 - 1)$
   (b) $F^* > F(1-\alpha; n_1 - 1, n_2 - 1)$
   (c) $F^* < F(\alpha; n_1 - 1, n_2 - 1)$

Proof that $s_1^2/s_2^2 \sim F(n_1 - 1, n_2 - 1)$ Under H_0

Under H_0, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and

Examples

Example 11-1. The average length of time for students to register for fall classes has been 35 minutes, with a standard deviation of 10 minutes. A new, web-based system was recently implemented. If a random sample of 12 students had an average registration time of 29 minutes with a standard deviation of 11.9 minutes under the new system, test the hypothesis that the population mean is now less than 35 minutes, using a .01 level of significance. Assume registration times are normally distributed.

Example 11.1 Details

Example 11.2

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a standard deviation of 4, while the samples of material 2 gave an average of 81 and a standard deviation of 5. Test the hypothesis that the two types of material exhibit the same variance in abrasive wear at the 0.10 level of significance. Assume that wear is normally distributed and that the variances in the two populations are equal.

Example 11.2 Details

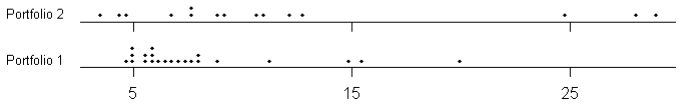
Example 11.3

The table below gives stock returns for two portfolios. Plot the data, and test for equality of variances and means using a .05 level of significance, assuming normality of returns.

Portfolio 1	Portfolio 2
8.05	9.09
6.2	10.93
4.93	12.2
20.09	4.54
5.85	8.89
5.77	3.4
4.88	10.54
5.02	24.79
5.51	7.53
6.57	6.65
7.95	7.55
14.99	12.68
7.54	28.05
7.03	28.9
5.5	4.48
8.94	
5.97	
15.58	
11.1	
6.62	
4.54	
7.39	

Dot Plots of Data

Dotplot for Portfolio 1-Portfolio 2

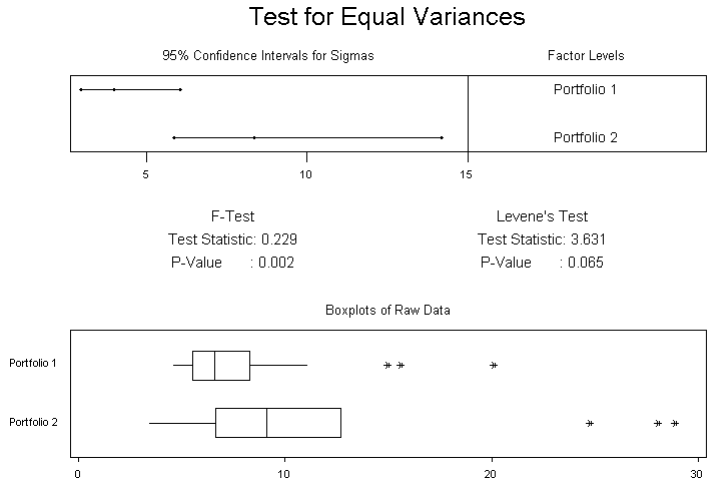


Example 11-3 via Minitab

Use: Stat >> Basic Statistics >> 2 Variances.

```
Level1     Portfolio 1
Level2     Portfolio 2
ConfLvl    95.0000

Bonferroni confidence intervals for standard deviations

  Lower     Sigma     Upper     N  Factor Levels

 2.97754   4.01132    6.0585    22  Portfolio 1
 5.88160   8.38027   14.2041    15  Portfolio 2


F-Test (normal distribution)


Test Statistic: 0.229
P-Value       : 0.002


Levene's Test (any continuous distribution)


Test Statistic: 3.631
P-Value       : 0.065
```


Example 11-3 via Minitab (cont)



Example 11-3 via Minitab (cont)

Stat >> Basic Statistics >> 2-sample t
(Do not specify equal variances)

```
Two-sample T for Portfolio 1 vs Portfolio 2

           N      Mean     StDev   SE Mean
Portfoli  22      8.00      4.01      0.86
Portfoli  15     12.01      8.38       2.2

Difference = mu Portfolio 1 - mu Portfolio 2
Estimate for difference:  -4.01
95% CI for difference: (-8.90, 0.88)
T-Test of difference = 0 (vs not =): T-Value = -1.72  P-Value = 0.102  DF = 18
```

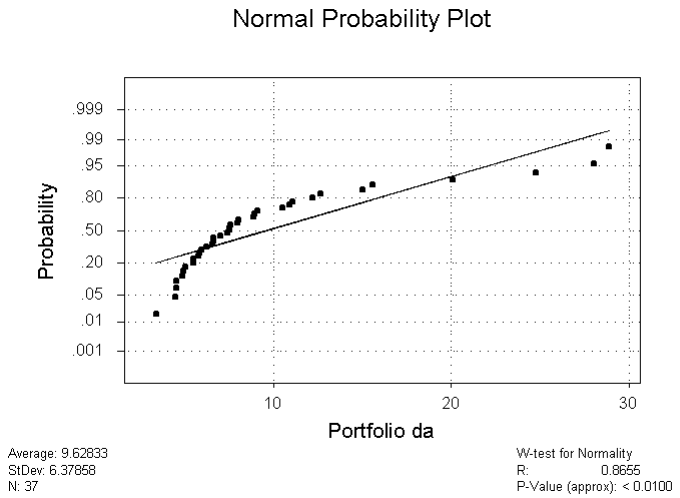

Testing for Normality via Minitab

Null hypothesis : Data are normal.
Alternative: Data not normal

Use Stat >> Basic Statistics >> Normality Test
(Use Ryan-Joiner option)

In the normal probability plot that accompanies the test, look to see if the data fall on a straight line. If not, this is an indication that the data do not follow a normal distribution.

Example 11-5 Test for Normality in the Combined Portfolio Data (stack columns)



What if the Data are not Normal?

Two options: 1) Transform it (take logs, square-roots, etc.)
2) Bootstrap it

1. Box-Cox Power Transformation.
Suppose X^{λ} is normally distributed for some λ:
- $\lambda = -1$ implies an inverse transformation needed
- $\lambda = .5$ implies a square-root transformation
- $\lambda = 0$ implies a log transformation (by definition)

We can add λ as a parameter and estimate it using normal-theory maximum likelihood!

Box-Cox via Minitab

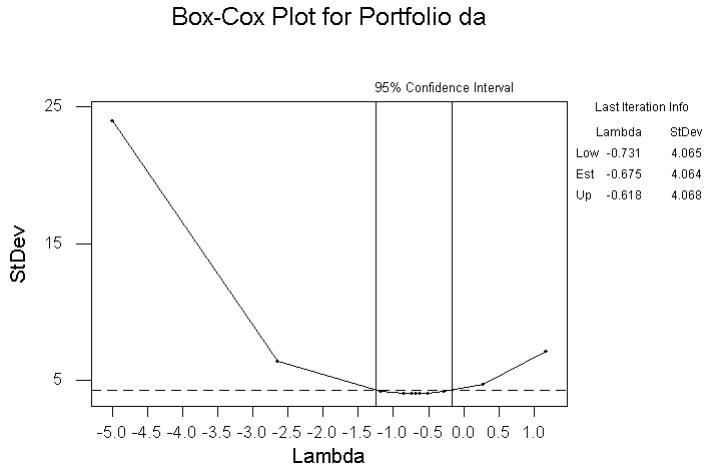
Stat > Control Charts >> Box-Cox Transformation

Put the data column in the "single column" box, the sample size in the "subgroup size" box and indicate the new column in the transformed data box.

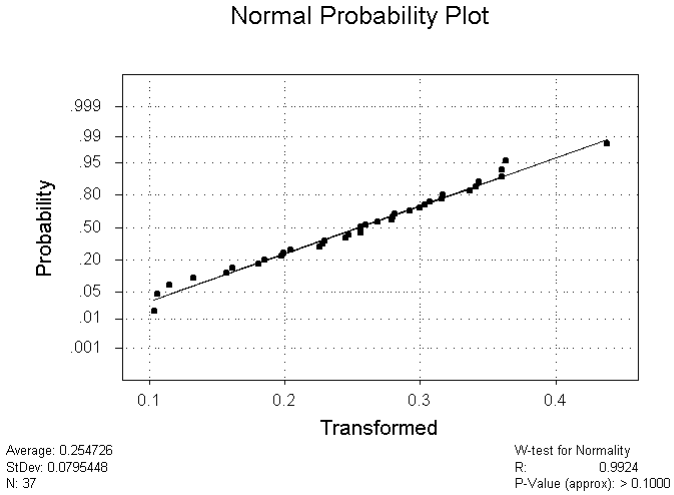
Read the power transformation parameter from "Est" in the plot legend that appears.

Caveat: Only works for positive data!!!

Example 11-4 Find Box-Cox for Portfolio Data (Combined Data)



Example 11-4 Results



2) Testing Based on the Bootstrap

Step 1: Compute the (1 - α)100% non-parametric bootstrap confidence interval for the parameter being tested.

Step 2: If the hypothesized value does not fall within the limits, reject H_0.

Note: This works only for two-sided alternatives. Recall that a two-sided hypothesis test will reject the null hypotheses if and only if the hypothesized value falls in the confidence interval

Chapter 15: Bayesian Inference

15.2: Decision Theory: We'll consider a 7-step process

Example 15.1-1 Sampling Inspection

A manufacturer receives a shipment of N items. A random sample of n items are inspected for defects. The supplier has guaranteed that the true proportion of defectives, p, is less than 5%. The manufacturer can either accept or reject the shipment.

1) Actions: a_1 = accept; a_2 = reject

2) True (unknown) states of nature: p

3) Data is $X = \{x_1, x_2, \ldots, x_n\}$; random sample of n items

Example 15.1-1 (continued)

5) Loss function: $l(p, d_i(X))$

	Decision $d_i(X)$	
State of Nature, p	Accept if $p \leq p_i$	Reject if $p > p_i$
$p < p_i$	\$0	\$2,000
$p \geq p_i$	\$5,000	\$0

6) Risk. Risk is the expected loss for each decision rule:

$$R(p, d_i) = E_{X|p}[l(p, d_i(X))]$$

7) Criteria. Since risk depends on p (the true state of nature), the best decision rule may depend on the true p. What do we do (we don't know p)??????

Decision Criteria

7.1 MINIMAX RULE. (Play it conservative!!!)

Choose d_i to minimize he worst-case value of p:

$$\min_{d_i}\{\max_{p\in[0,1]} R(p, d_i)\}$$

Problem: Worst-case value may be p = 1; but you don't really believe p = 1 is very likely.

7.2 BAYES' RULE: Treat p as a random variable. Express your prior beliefs about p in the form of a prior probability distribution. Then choose the decision rule d_i that minimizes expected risk:

$$\min_{d_i}\{E_p[R(p, d_i)]\} = \min_{d_i} B(d_i)$$

where $B(d_i)$ is the *Bayes' risk* of d_i.

Example 15.2-2 Estimation

Objective is to estimate a parameter (say) μ on the basis of a sample $X = \{x_1, x_2, \ldots, x_n\}$

1) **Action space: Estimate μ with a value d(X), $-\infty < d(X) < \infty$.**
2) **States of nature: $-\infty < \mu < \infty$**
3) **Data: $X = \{x_1, x_2, \ldots, x_n\}$**
4) **Decision rule: choose an appropriate function (estimator) d(X).**
5) **Loss function: We will use the *quadratic loss* function:**

$$l(\mu, d(X)) = [\mu - d(X)]^2$$

Example 15.2-2 Estimation (continued)

6) **Risk:**

$$R(\mu, d(X)) = E_{X|\mu}[(\mu - d(X))^2]$$
$$= MSE[\mu, d(X)]$$

7) **Criteria:**

7.1. Minimax risk: Choose d to minimize (over μ) the maximum quadratic loss.

7.2 Bayes (expected) risk: Express your prior beliefs about m in the form of a prior distribution g(μ). Then:

$$\min_d \{E_\mu[R(\mu, d)]\} = \min_d B(d)$$
$$= \min_d \int_{-\infty}^{\infty} MSE[\mu, d(X)] g(\mu) d\mu$$

Example 15.2-3 Foundation Construction

As part of a foundation of a building, a steel section is to be driven down to a firm stratum below ground.

1. **Actions:** a_1: **Choose a 40-ft section**
   a_2: **Choose a 50-ft section**

2. **States of nature:** θ_1: **True depth is 40 ft.**
   θ_2: **True depth is 50 ft.**

3. **Data: A depth sounding is taken by means of a sonic test; result is X (one observation).**

Example 15.2-3 (continued)

Distribution of X|θ is:

X	P(X|θ=40)	P(X|θ=50)
40	0.6	0.1
45	0.3	0.2
50	0.1	0.7

4) Decision rules:

$d_i(X)$	X = 40	X = 45	X = 50
$d_1(X)$	a_1 (40 ft)	a_1 (40 ft)	a_1 (40 ft)
$d_2(X)$	a_1 (40 ft)	a_2 (50 ft)	a_2 (50 ft)
$d_3(X)$	a_1 (40 ft)	a_1 (40 ft)	a_2 (50 ft)
$d_4(X)$	a_2 (50 ft)	a_2 (50 ft)	a_3 (50 ft)

Example 15.2-3 (continued)

5) Loss function:

	State of Nature	
Action	θ_1: Depth = 40	θ_2: Depth = 50
a_1	\$0	\$400
a_2	\$100	\$0

	$l(\theta_1, d_i(X))$			$l(\theta_2, d_i(X))$		
$d_i(X)$	X = 40	X = 45	X = 50	X = 40	X = 45	X = 50
$d_1(X)$	\$0	\$0	\$0	\$400	\$400	\$400
$d_2(X)$	\$0	\$100	\$100	\$500	\$0	\$0
$d_3(X)$	\$0	\$0	\$100	\$500	\$0	\$0
$d_4(X)$	\$100	\$100	\$100	\$0	\$0	\$0

Example 15.2-3 (continued)

5) **Risk. For θ_1 (40 feet true depth):**

$$R(\theta_1, d_1) = .6 \times 0 + .3 \times 0 + .1 \times 0 = 0$$
$$R(\theta_1, d_2) = .6 \times 0 + .3 \times 100 + .1 \times 100 = 40$$
$$R(\theta_1, d_3) = .6 \times 0 + .3 \times 0 + .1 \times 100 = 10$$
$$R(\theta_1, d_4) = .6 \times 100 + .3 \times 100 + .1 \times 100 = 100$$

For θ_2 (50 feet true depth):

$$R(\theta_2, d_1) = .1 \times 400 + .2 \times 400 + .7 \times 400 = 400$$
$$R(\theta_2, d_2) = .1 \times 400 + .2 \times 0 + .7 \times 0 = 40$$
$$R(\theta_2, d_3) = .1 \times 400 + .2 \times 400 + .7 \times 0 = 120$$
$$R(\theta_2, d_4) = .1 \times 400 + .2 \times 400 + .7 \times 400 = 0$$

Example 15.2-3 (continued)

7) **Criterion:**

7.1 Minimax Rule: d_2

7.2 Bayes Rule (Bayes Risk): Suppose from geological data, we believe:

$g(\theta_1) = .8$ and $g(\theta_2) = .2$

$$B(d_1) = E_\theta R(\theta, d_1) = .8 \times 0 + .2 \times 400 = 80$$
$$B(d_2) = .8 \times 40 + .2 \times 40 = 40$$
$$B(d_3) = .8 \times 10 + .2 \times 120 = 32$$
$$B(d_4) = .8 \times 100 + .2 \times 0 = 80$$

Therefore the Bayes' Rule is d_3

Example 15.2-4

A manufacturer produces items in lots of 21. One item is selected at random and tested.

If the item is defective:

- **Can sell the remaining 20 items for $1 each with a double-your-money-back guarantee on each item, or**
- **Junk the whole lot for a total cost of $1**

If the item is not defective:

- **You will sell the remaining 20 items as above**

Find the minimax and Bayes' rules.

Example 15.2-4 (continued)

1. **Actions: a_1 = sell; a_2 = junk**
2. **States of nature: $0 \le k \le 21$ defects**
3. **Data: X = 0 if defective; X = 1 if not defective**
4. **Decision rules:**
   - **$d_1(X)$: Sell if X = 1, junk if X = 0**
   - **$d_2(X)$: Sell in either case**
5. **Loss function:**

$$l(k, d_1(X)) = \begin{cases} -20 + 2k & \text{if } X = 1 \\ 1 & \text{if } X = 0 \end{cases}$$

$$l(k, d_2(X)) = \begin{cases} -20 + 2k & \text{if } X = 1 \\ -20 + 2(k-1) & \text{if } X = 0 \end{cases}$$

Example 15.2-4 (continued)

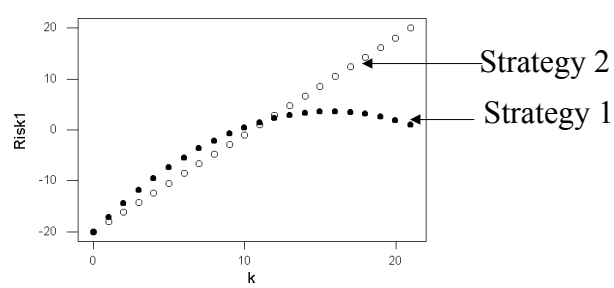
6. **Risk function:**

$$
\begin{aligned}
R(k, d_1) &= E_{X|k} l(k, d_1(X)) \\
&= (-20 + 2k) \times \left(1 - \frac{k}{21}\right) + (1) \times \frac{k}{21} \\
&= -20 + 3k - \frac{2k^2}{21} \\
R(k, d_2) &= (-20 + 2k) \times \left(1 - \frac{k}{21}\right) + [-20 + 2(k-1)] \times \frac{k}{21} \\
&= -20 + \frac{40k}{21}
\end{aligned}
$$

Example 15.2-4 (continued)

7. **Criteria:**

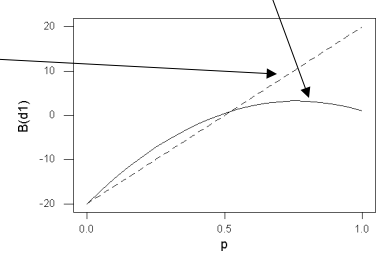
7.1 Minimax criterion: d_1 (Strategy 1)



Example 15.2-4 (continued)

7.2 Bayes' Rule: Assume k is Binomial(21, p). Then:

$$B(d_1) = E_k[-20 + 3k - \frac{2k^2}{21}]$$
$$= -20 + 3E[k] - \frac{2E[k^2]}{21}$$
$$= -20 + 3(21p) - \frac{2}{21}[21p(1-p) + (21p^2)]$$
$$B(d_2) = E_k[-20 + \frac{40k}{21}]$$
$$= -20 + \frac{40(21p)}{21}$$



d_1: best for p > .52

d_2: best for p < .52

15.2.2 Posterior Analysis

- **This has been a lot of work--we need a short cut!**
- **Forget minimax criterion; we'll focus on Bayes' risk**

Notation:

1) **The prior density or probability mass function for θ: g(θ)**
2) **The conditional probability density or probability mass function of X| θ: f(x| θ)**
3) **The joint distribution of X and θ:**

$$f_{X,\theta}(x, \theta) = f(x|\ \theta)\ g(\theta)$$

Posterior Analysis: Setup

4) **Marginal distribution of X:**

$$f_X(x) = \begin{cases} \int_\theta f(x|\theta)g(\theta)d\theta & \theta \text{ continuous} \\ \sum f(x|\theta_i)g(\theta_i) & \theta \text{ discrete} \end{cases}$$

5) **We can apply Bayes' theorem to find the *posterior distribution* of θ:**

$$\begin{aligned} h(\theta|X) &= \frac{f_{X,\theta}(X,\theta)}{f_X(X)} \\ &= \frac{f(X|\theta)g(\theta)}{\int_\theta f(X|\theta)g(\theta)d\theta} \\ &= \frac{lik(\theta|X)g(\theta)}{\int_\theta lik(\theta|X)g(\theta)d\theta} \end{aligned}$$

"Posterior" because it is after observing the data, X

Posterior Risk: Short cut to finding the Bayes Rule

Definition: The *posterior risk* of d is the expected loss, where the expecation is taken with respect to the posterior distribution of θ:

$$E_\theta l(\theta, d(X)) = \int_\theta l(\theta, d(X))h(\theta|X)d\theta$$

Theorem: (the short cut)

Suppose there is a function $d_0(X)$ that minimizes the posterior risk. Then d_0 is a Bayes' rule

Posterior Risk: Short cut to finding the Bayes Rule

Proof:

$$
\begin{aligned}
B(d) &= E_\theta[R(\theta, d)] \\
&= E_\theta[E_{X|\theta} l(\theta, d(x))] \\
&= \int \left[\int l(\theta, d(x)) f(x|\theta) dx \right] g(\theta) d\theta \\
&= \int \int l(\theta, d(x)) f(x|\theta) dx g(\theta) d\theta \\
&= \int \int l(\theta, d(x)) f_{X,\theta}(X, \theta) dx d\theta \\
&= \int \int l(\theta, d(x)) h(\theta|x) f_X(x) dx d\theta \\
&= \int \left[\int l(\theta, d(x)) h(\theta|x) d\theta \right] f_X(x) dx \\
&\geq \int d_0(X) f_X(x) dx
\end{aligned}
$$

So d_0 minimizes B(d).

Result: Using Posterior to Find Bayes' Rule

Step 1. **Calculate the posterior distribution h(θ|X).**

Step 2. **For each possible action *a*, compute the posterior risk:**

$$\int l(\theta, a) h(\theta|X) d\theta$$

Step 3. **The action *a** that minimizes the posterior risk is the Bayes' rule.**

Summary: **Choose the action (in statistics, the estimator) that minimizes the posterior risk.**

Example 15.2-5 Foundation Construction, Again

Step 1 Assume now that we observed X = 45. The posterior distribution of θ is:

$$h(\theta_1|X=45) = \frac{f(45|\theta_1)g(\theta_1)}{\sum_{i=1}^{i=2} f(45|\theta_i)g(\theta_i)}$$
$$= \frac{.3 \times .8}{.3 \times .8 + .2 \times .2} = .86$$

$$h(\theta_2|X=45) = 1 - .86 = .14$$

Step 2 Posterior expected loss (posterior risk):

$$\sum_{i=1}^{2} l(\theta_i, a_1)h(\theta_i|45) = 0(.86) + 400(.14) = 56$$

$$\sum_{i=1}^{2} l(\theta_i, a_2)h(\theta_i|45) = 100(.86) + 0(.14) = 86$$

Step 3 Action a_1 is the Bayes' rule.

15.2.3 Classification and Hypothesis Testing

Suppose the population consists of m classes (or there are m distinct populations). You observe X = $\{x_1, x_2, \ldots, x_n\}$. To which class does X belong?

Examples

1) Voting ballot optical scanners (pattern recognition). A observations on a set of pixels $\{x_1, x_2, \ldots, x_n\}$ are either zero or one depending on whether or not ink is present. Does the patter suggest a vote for candidate 1 or candidate 2?

2) Medical diagnoses. There are n physiological measurements X. What are the posterior probabilities of the various possible diseases 1, …, m.

Examples, Continued

3) Palm pilot (pattern recognition again). X = pixel readings again. What letter, A, B, …, C is indicated?

4) Voice recognition (pattern recognition again). X = {audio pattern}. What word does the pattern indicate?

In all cases we need:

- Class memberships: $\theta_1, \ldots, \theta_m$
- Prior probabilities: $g(\theta_1), \ldots, g(\theta_m)$
- Data: $X = \{x_1, x_2, \ldots, x_n\}$
- Distribution of data given class membership $f(X|\theta)$

Choosing the Class via Bayes Rule

Given all that, the posterior distribution of θ is:

$$\begin{aligned} h(\theta|X) &= P(\Theta = \theta_k|X) \\ &= \frac{f(X|\theta_k)g(\theta_k)}{\sum_{j=1}^{m} f(X|\theta_j)g(\theta_j)} \end{aligned}$$

Suppose l_{ij} is the loss for choosing class j when class i is the correct class. Then the posterior risk of choosing class i is:

$$\begin{aligned} B(d_i) &= \sum_{k=1}^{m} l_{ki} h(\theta_k|X) \\ &= \sum_{k=1}^{m} l_{ik} \left[\frac{f(X|\theta_k)g(\theta_k)}{\sum_{j=1}^{m} f(X|\theta_j)g(\theta_j)} \right] \end{aligned}$$

Bayes rule: Choose class that minimizes $B(d_i)$

Special Case: Zero-One Loss

Suppose the loss function is:

$$l_{ij} = \begin{cases} 0; \text{ if } i = j \text{ (we're right)} \\ 1; \text{ if } i \neq j \text{ (we're wrong)} \end{cases}$$

Then the posterior expected loss for choosing class *i* is:

$$\begin{aligned} E(l, i|X) &= \sum_{j=1}^{m} l(\theta_j, i) h(\theta_j|X) \\ &= \sum_{j \neq i} h(\theta_j|X) \\ &= 1 - h(\theta_i|X) \\ &= \text{posterior probability of misclassification} \end{aligned}$$

Choosing the Class via Bayes Rule (continued)

The posterior risk of choosing class *i* is:

$$B(d_i) = E_\theta[1 - h(\theta_i|X)] = 1 - h(\theta_i|X)$$

Bayes rule: Minimize posterior risk by choosing the class *i* that maximizes the posterior probability!

Hypothesis Testing: Bayesian Perspective

Bayesian perspective on hypothesis testing is simple:

Hypothesis testing is silly:

- **Type I and Type II errors are irrelevant because θ is a random variable.**
- **You simply need to specify your loss function and chose the decision (accept or reject) that minimizes posterior risk.**

15.2.5 Bayesian Estimation

Recall Example 15.2-2. We wish to estimate θ using a point estimate d(X) = $\hat{\theta}$, based on squared-error loss:

$$l(\theta, \hat{\theta}) = [\theta - \hat{\theta}]^2$$

Step 1. Calculate the posterior distribution, h(θ |X).

Step 2. Calculate the posterior risk:

$$E[l(\theta, \hat{\theta})|X] = \int_\theta (\theta - \hat{\theta})^2 h(\theta|X) d\theta = E_\theta[(\theta - \hat{\theta})^2|X]$$

Step 3. Find the value of $\hat{\theta}$ that minimizes posterior risk:

$$\begin{aligned} E_\theta[(\theta - \hat{\theta})^2|X] &= Var_\theta[(\theta - \hat{\theta})^2|X] + [E_\theta(\theta - \hat{\theta}|X)]^2 \\ &= Var_\theta(\theta) + [E_\theta(\theta) - E_\theta(\hat{\theta})]^2 \\ &= Var_\theta(\theta) + [E_\theta(\theta) - \hat{\theta}]^2 \end{aligned}$$

which is minimized by $\theta = E_\theta(\theta)$, the posterior mean of θ.

Bayesian Estimation of θ: Summary

Step 1. Give a prior distribution for θ.

Step 2. Employ a squared-error loss function.

Step 3. Collect data and determine f(X|θ) = lik(θ|X)

Step 4. Calculate the posterior distribution of θ:

$$h(\theta|X) = \frac{lik(\theta|X)g(\theta)}{\int_\theta lik(\theta|X)g(\theta)}$$

Step 5. Use: $\hat{\theta} = E_\theta(\theta)$, the mean of the posterior distribution of θ for a point estimate.

Step 6. (1 - α)100% Bayesian confidence intervals, called *credibility intervals*, can be obtained from the α/2 and 1 - α /2 percentiles of the posterior distribution of θ.

Example 15.2-6 Bayes Estimate of p

A single coin is flipped. Estimate p, the probability of heads from the outcome. Assume the coin may be biased.

1) Prior distribution: p ~ Uniform[0, 1]; g(p) = 1, 0 ≤ p ≤ 1
2) We'll assume squared-error loss.
3) X = 1 (coin came up heads)
4) Posterior distribution of p:

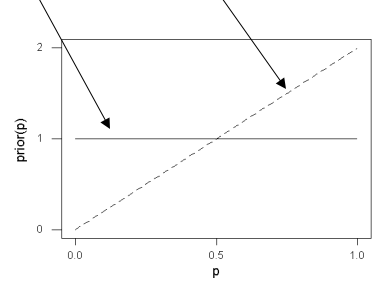
$$h(p|X = 1) = \frac{f(1|p)g(p)}{\int_0^1 f(1|p)g(p)dp}$$
$$= \frac{p \times 1}{\int_0^1 pdp} = 2p$$

Example 15.2-6 (Continued)

5) **Bayes estimate is the mean of the posterior distribution:**

$$\hat{p} = \int_0^1 p(2p)\,dp$$
$$= \frac{2}{3}$$

Picture of prior and posterior distribution of θ:



Example 15.2-6 (Continued)

6) **90% credibility interval (recall calculating percentiles):**

$$F(p^*) = \int_0^{p^*} 2p\,dp = 2\left[\frac{p^2}{2}\right]_0^{p^*} = (p^*)^2$$

5.0 percentile of posterior:

$$.05 = [p^*]^2 \Rightarrow p^* = \sqrt{.05} = .2236$$

95.0 percentile of posterior:

$$.95 = [p^*]^2 \Rightarrow p^* = \sqrt{.95} = .9747$$

90% credibility interval: (.2236, .9747)

Example 15.2-6 (Continued)

Question 1: What would be the frequentist estimate of p?

Question 2: What would the 95% confidence interval be?

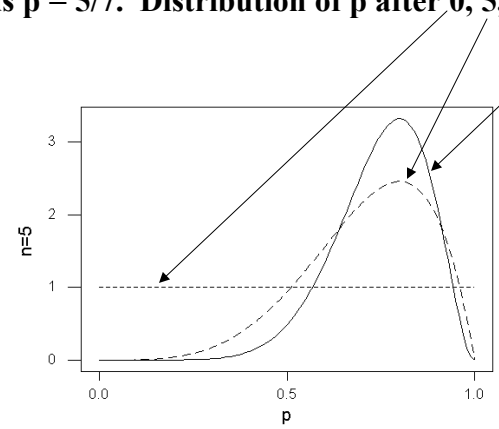
Example 15.2-7 Binomial Case

1) **Prior distribution: p ~ Uniform[0, 1]; g(p) = 1, 0 ≤ p ≤ 1**
2) **We'll assume squared-error loss.**
3) **X = 8 (8 heads, 2 tails)**
4) **Posterior distribution of p:**

$$
\begin{aligned}
h(p|X=8) &= \frac{f(8|p)g(p)}{\int_0^1 f(8|p)g(p)dp} \\
&= \frac{\binom{10}{8}p^8(1-p)^2 \times 1}{\int_0^1 \binom{10}{8}p^8(1-p)^2 dp} \\
&= \frac{1}{.0020202}p^8(1-p)^2
\end{aligned}
$$

Example 15.2-7 Binomial Case

5) **Expected value of the posterior distribution is $\hat{p}$ = .75. Note: same calculation for X = 4 heads in n = 5 flips yields $\hat{p}$ = 5/7. Distribution of p after 0, 5, and 10 flips:**



Bayesian Inference for the Normal Distribution

We want to give a Bayesian estimate of μ. We will assume the data are normal and we will assume a normal prior for μ.

1) **Assume $\mu \sim N(\mu_0, \sigma_0^2)$**
2) **We will assume squared-error loss as before**
3) **We have a sample of size 1: $X \sim N(\mu, \sigma^2)$**
4) **Posterior distribution of μ:**

$$f(x|\mu)g(\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right] \times \frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$
$$\propto exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2 + \frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$
$$= exp\frac{-1}{2}\left[\mu^2\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + \left(\frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right)\right]$$

Normal Posterior, Continued

$$
\begin{aligned}
&= exp\left[\frac{-1}{2}[a\mu^2 - 2\mu b + c]\right] \\
&= exp\left[\frac{-a}{2}\left[\mu^2 - 2\frac{2b}{a} + \frac{c}{a}\right]\right] \\
&= exp\left[\frac{-a}{2}\left[\left(\mu - \frac{b}{a}\right)^2 + \frac{c}{a} - \frac{b^2}{a^2}\right]\right] \\
&= exp\left[\frac{-a}{2}\left(\mu - \frac{b}{a}\right)^2\right] \times exp\left[\frac{-a}{2}\left[\frac{c}{a} - \left(\frac{b}{a}\right)^2\right]\right]
\end{aligned}
$$

Therefore the posterior distribution of μ is

$$h(\mu|X) \propto exp\left[\frac{-1}{2}\frac{\left(\mu - \frac{b}{a}\right)^2}{\frac{1}{a}}\right]$$

Normal Posterior, Continued

which is normal with mean:

$$
\begin{aligned}
\hat{\mu} &= \frac{b}{a} = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}} \\
&= w_1 x + w_2 \mu_0
\end{aligned}
$$

$$w_1 = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}} \qquad w_2 = \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

The posterior variance is:

$$\frac{1}{a} = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Normal Posterior, Continued

So the posterior mean is a weighted average of the sample observation and the prior mean μ_0, where the weights are inversely proportional to the respective variances. For a sample of size n, a similar analysis (see text) leads to:

$$\hat{\mu} = \frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$= w_1\bar{X} + w_2\mu_0$$

$$w_1 = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \qquad w_2 = \frac{\frac{1}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\text{Posterior variance} = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Example 15.3-1 Sequential updates to posterior

Prior distribution of μ is N(2, 4) and the X_i are N(4, 1)

Stage 1: **We observe $X_1 = 3.59$, $X_2 = 5.52$ ($\bar{X} = 4.55$)**

$$\hat{\mu} = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\bar{X} + \frac{\frac{1}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu_0$$

$$= .89\bar{X} + .11\mu_0$$

$$= 4.27$$

Stage 2: **We observe $X_3 = 3.93$, $X_4 = 4.71$ ($\bar{X}_4 = 4.44$)**

$$\hat{\mu} = .94\bar{X} + .06\mu_0$$

$$= 4.30$$

Example 15.3-1 (continued)

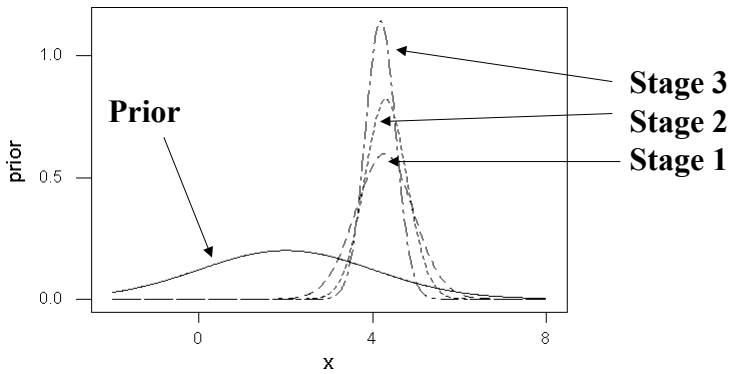
Prior distribution of μ is N(2, 4) and the X_i are N(4, 1)

Stage 3: $X_5 = 4.40$, $X_6 = 5.06$, $X_7 = 3.68$, $X_8 = 3.14$, so $\bar{X} = 4.25$:

$$\hat{\mu} = .97\,\bar{X} + .03\,\mu_0$$
$$= 4.18$$

Pictures of the four stages of the posterior distribution follow:

Example 15.3-1 (continued)



Notes

1. **If we let the variance of the normal prior go to infinity, we get a uniform "distribution" on the real number line. This is the idea behind a "vague" or "noninformative" prior. Basically, we sometimes use a uniform distribution on the set of possible values for the parameter, in an effort to minimize the influence of the prior.**

2. **In such a case, the posterior distribution is the same as the likelihood!**

3. **Should be apparent: Bayesian methods usually require a lot of numerical computing to get estimates, intervals.**

15.3 The Subjectivist Point of View

What is probability (of say, heads)?

Frequentist: Long-run frequency from flipping the coin.

Bayesian: Suppose a game involves one flip. The payout is $X if heads. Then p(heads) is the ratio of the amount ($Y) the Bayesian is willing to pay to play the game, to the cost of the game $X. So if the game costs $1 and the Bayesian is willing to pay $.50, then for the Bayesian:

P(heads) = $.50/$1.00 = .5

Bayesian probability is personal and may vary from person to person