

# Economics 623: Econometrics \*

Matthew Chesnes

Updated: January 1, 2005

---

\*These are Matthew Chesnes' notes from a course taught by Ingmar Prucha.

# 1 Lecture 1: August 31, 2004

## 1.1 Motivation for the Course

- Consider a simple example of a linear consumption function:

$$c = a + by.$$

Where  $c$  is consumption,  $y$  is income, and  $b$  is the Marginal Propensity to Consume (MPC). Normally we would start with data on  $c$  on  $y$  and then fit a line to the data. This line would just be the average relationship between consumption and income so the real relationship is:

$$c_i = a + by_i + u_i,$$

where  $u_i$  is some random observation or disturbance term. We might assume the following:

$$E[u_i] = 0,$$

$$E[u_i^2] = \sigma^2,$$

$u_i$ 's are independent.

In doing all of this, we have used Probability, the idea of a Random Variable (RV), a Distribution, the concept of Independence, and Expectation. Hence we have the motivation for the course.

## 1.2 Probability

- Consider a coin toss, head or tail. The sample space is defined as:

$$\Omega = \{H, T\}.$$

The set of possible events (or the power set):

$$\left\{ \{H\}, \{T\}, \Omega, \emptyset \right\}.$$

Note the probabilities of each event:

$$Pr(H) = 0.5, \quad Pr(T) = 0.5, \quad Pr(\Omega) = 1, \quad Pr(\emptyset) = 0.$$

- Thus we have a probability function defined on the subsets of the sample space,  $\Omega$ .

## 1.3 Random Variable

- Consider the mapping:

$$X : \Omega \mapsto \{0, 1\}.$$

$$X(H) = 1, \quad X(T) = 0.$$

This is just a more convenient notation than heads and tails say.  $X$  is a RV.

## 1.4 Distribution Function

- Consider:

$$F(x) = Pr(X \leq x).$$

This is the (cumulative) distribution function and produces values between 0 and 1.

## 1.5 Independence

- $A$  and  $B$  are independent if the knowledge of  $A$  happening does not effect the probability of  $B$  happening.

## 1.6 Expectation

- Average, etc.

## 1.7 Returning to the Model

- Consider again the model:

$$c_i = a + by_i + u_i.$$

- We would like an estimate of  $b$ . One might be the Ordinary Least Squares estimator (OLS):

$$\hat{b} = \frac{\sum(y_i - \bar{y})(c_i - \bar{c})}{\sum(y_i - \bar{y})^2}.$$

We could also write:

$$\hat{b} = b + \underbrace{a_1u_1 + a_2u_2 + \cdots + a_nu_n}_{\text{Disturbances—Random Variables}}.$$

So  $\hat{b}$  is a random variable and it has some distribution.

- Now we need to compute the distribution of functions of RVs.  $\hat{b}$  might be our point estimator. We might want to construct a confidence interval (Interval Estimation), or we might want to check a hypothesis (Hypothesis Testing).
- If there is time, at the end of the course, we will study asymptotic theory.

## 2 Lecture 2: September 2, 2004

### 2.1 Probability

#### Random Experiments, Sample Spaces, and Event Spaces

- Random Experiment: Unknown outcome, repeatable, and it should describe a collection of possible outcomes.
- Sample Space: The collection of all possible outcomes,  $\Omega$ .
- Sample Point: Possible outcome,  $\omega \in \Omega$ .
- Event: A subset  $A \subset \Omega$ . An event occurs if the outcome of the experiment belongs to  $A$ , ie,  $\omega \in A$ .
- Any single events that contain just one of the elements of  $\Omega$  are called basic events or singletons.
- In general, if there are 8 outcomes (8 elements in the set  $\Omega$ ), then there should be  $2^8$  events (which are all possible combination of outcomes.)
- So for a single coin flip:

$$\Omega = \{\{H\}, \{T\}\}.$$

And the  $2^2 = 4$  events:

$$\{H\}, \{T\}, \emptyset, \Omega.$$

### 2.2 Set Theory - Covered in Discussion

### 2.3 Review of the definition of a function

- **Definition:** A function is a rule in which elements of a set  $A$  are associated with the elements of a set  $B$ .  $f : A \mapsto B$ .
- So for  $A = [0, 1]$ ,  $B = [0, \infty]$ ,  $f(x) = x^2 \forall x \in A$  is monotonically increasing.
- So for  $A = [-1, 1]$ ,  $B = [0, \infty]$ ,  $f(x) = x^2 \forall x \in A$  is NOT monotonically increasing. So the key to all of this is that to have a rigerous definition of a function, you must include the domain space and the counter-domain because a function will have different properties over different domains.
- Example. Let  $\mathfrak{A} = \{A : A = (a, b] \subseteq \mathfrak{R}\}$ .  $f : \mathfrak{A} \mapsto \mathbb{N}$  with  $f(A)$  equal to the number of integers in  $A$ .  $f$  is a “set function” in that the domain is a subset of some other space.

## Definition of Probability

- Classical definition. Consider the sample space:

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}.$$

And an arbitrary event:

$$A = \{w_{i1}, w_{i2}, \dots, w_{in_A}\}.$$

So there are  $n_A$  elements in  $A$ . Then if all the outcomes in  $\Omega$  are equally likely,

$$P(A) = \frac{n_A}{n}.$$

- Now consider the concept of relative frequency. Suppose we have  $\Omega$ ,  $A$ . Define the absolute frequency as:

$$f_n(A),$$

or the frequency which  $A$  occurs among  $n$  trials. Then we have the RELATIVE frequency:

$$P_n(A) = \frac{f_n(A)}{n}.$$

- Consider the properties of  $f_n(A)$ :
  - 1)  $0 \leq f_n(A) \leq n$ .
  - 2)  $f_n(\emptyset) = 0$ ,  $f_n(\Omega) = n$ .
  - 3) Suppose  $A_1, A_2, \dots \subset \Omega$  and  $A_i \cap A_j = \emptyset$  for  $i \neq j$ . (Disjoint) Then:

$$f_n(A_1 \cup A_2) = f_n(A_1) + f_n(A_2).$$

- Consider the properties of  $p_n(A)$ :
  - 1)  $0 \leq p_n(A) \leq 1$ .
  - 2)  $p_n(\emptyset) = 0$ ,  $p_n(\Omega) = 1$ .
  - 3) Suppose  $A_1, A_2, \dots \subset \Omega$  and  $A_i \cap A_j = \emptyset$  for  $i \neq j$ . (Disjoint) Then:

$$p_n(A_1 \cup A_2) = p_n(A_1) + p_n(A_2).$$

- The same is of course true for more than two sets of disjoint events. Note that the sets of events must be countable for this to work.
- Axiomatic definition of probability. Define a function:

$$p = \mathfrak{A} \mapsto [0, 1].$$

Where  $\mathfrak{A}$  is a set of events.

- **Definition:** Let  $\Omega$  be an arbitrary nonempty set. A set  $\mathfrak{A}$  of subsets of  $\Omega$  is called an Algebra or Field if:

- 1)  $\emptyset \in \mathfrak{A}, \Omega \in \mathfrak{A}$ .
- 2) If  $A \in \mathfrak{A} \implies A^c \in \mathfrak{A}$ .
- 3) If  $A_1, A_2, \dots, A_n \in \mathfrak{A} \implies \cup_{i=1}^n A_i \in \mathfrak{A}$ .

- **Definition:** Let  $\Omega$  be an arbitrary nonempty set. A set  $\mathfrak{A}$  of subsets of  $\Omega$  is called a  $\sigma$ -Algebra or  $\sigma$ -Field if:

- 1)  $\emptyset \in \mathfrak{A}, \Omega \in \mathfrak{A}$ .
- 2) If  $A \in \mathfrak{A} \implies A^c \in \mathfrak{A}$ .
- 3) If  $A_1, A_2, \dots \in \mathfrak{A} \implies \cup_{i=1}^{\infty} A_i \in \mathfrak{A}$ .
- Hence this definition is only relevant if  $\Omega$  is infinite. The pair  $(\Omega, \mathfrak{A})$ , is called a measurable space. Any set  $A \in \mathfrak{A}$  is called a  $\mathfrak{A}$ -measurable set.

- Note that every  $\sigma$ -algebra is also an algebra.
- If  $\mathfrak{A}$  is a  $\sigma$ -algebra, then it follows from DeMorgan's Law that:

$$A_1, A_2, \dots \in \mathfrak{A} \implies \cap_{i=1}^{\infty} A_i \in \mathfrak{A}.$$

- Let  $\mathfrak{A}$  be a  $\sigma$ -algebra in  $\Omega$ . Then for any  $B \subseteq \Omega$ ,

$$B \cap \mathfrak{A} = \{B \cap A : A \in \mathfrak{A}\}$$

is a  $\sigma$ -algebra and it is called the TRACE of  $\mathfrak{A}$  in  $B$ . Consequently, a  $\sigma$ -algebra is closed under the formation of complements, countable unions, and countable intersections.

- Example.

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

$$A = \{1, 2\}.$$

Then:  $\mathfrak{A} = \{\emptyset, \Omega, A, A^c\}$  is an algebra.

### 3 Lecture 3: September 9, 2004

- **Definition:** Probability Measure and Probability Space. Consider an arbitrary non-empty set  $\Omega$  and a  $\sigma$ -algebra,  $\mathfrak{A} \in \Omega$ . A function,

$$P : \mathfrak{A} \mapsto [0, 1],$$

is called a probability measure (or probability function, or probability set function) if it satisfies the following conditions:

- $P(\emptyset) = 0, P(\Omega) = 1.$
- $P(A) \geq 0.$
- $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$  if  $A_i \cap A_j = \emptyset$  for  $i \neq j.$

The triple  $(\Omega, \mathfrak{A}, P)$  is called a probability space.

#### 3.1 Properties of the Probability Measure

- **Theorem 1:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space. If  $A \in \mathfrak{A} \Rightarrow P(A) = 1 - P(A^c).$   
Proof: Note  $\Omega = A \cup A^c.$   $P(\Omega) = 1 = P(A \cup A^c) = P(A) + P(A^c).$
- **Theorem 2:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space. Let  $A_1, A_2 \in \mathfrak{A}.$   $A_1 \subseteq A_2 \Rightarrow P(A_1) \leq P(A_2).$   
Proof:  $A_2 = A_1 \cup (A_2 - A_1).$   $P(A_2) = P(A_1) + P(A_2 - A_1) \geq P(A_1).$
- **Theorem 3:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space. For any  $A \in \mathfrak{A},$  we have  $0 \leq P(A) \leq 1.$   
Proof: Note  $A \subseteq \Omega \Rightarrow P(A) \leq P(\Omega) = 1.$   $\emptyset \subseteq A$  so  $0 = P(\emptyset) \leq P(A).$
- **Theorem 4:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space.  $A_1, A_2 \in \mathfrak{A}.$  Then  $P(A \cup B) = P(A) + P(B) - P(A \cap B).$   
Proof by picture (G-3.1)
- Note the probabilities are defined for EVENTS and NOT OUTCOMES. Consider a sample space:

$$\Omega = \{H, T\}.$$

An algebra:

$$\mathfrak{A} = \{\Omega, \emptyset, \{H\}, \{T\}\}.$$

We define probabilities for elements of  $\mathfrak{A},$  not the outcomes (elements of  $\Omega).$  So we have:

$$H \in \{H\}.$$

$\{H\} \in \mathfrak{A}$  Define Probability of this only.

$$H \notin \mathfrak{A}.$$

$$\{H\} \notin \Omega.$$

- **Theorem 5:** (Continuity of  $P$  from above and below). Let  $(\Omega, \mathfrak{A}, P)$  be a probability space.  $A_1, A_2, \dots \in \mathfrak{A}$  with  $A_{i+1} \subseteq A_i$ . So each subsequent set  $A$  is a subset of all previous. Then:

$$\lim_{i \rightarrow \infty} P(A_i) = P\left(\bigcap_{i=1}^{\infty} A_i\right).$$

The set  $A_i = [c - \frac{1}{i}, c + \frac{1}{i}]$  would be an example of this setup. If  $A_i \subseteq A_{i+1}$ , so the  $A$ 's are getting larger (containing all the previous sets), then:

$$\lim_{i \rightarrow \infty} P(A_i) = P\left(\bigcup_{i=1}^{\infty} A_i\right).$$

- So for example. Let  $\Omega = [0, 1]$  and  $0 \leq a < b \leq 1$  with  $P([a, b]) = b - a$ . For  $0 \leq c \leq 1$ , let:

$$A_i = [c - \frac{1}{i}, c + \frac{1}{i}] \cap [0, 1].$$

Thus  $A_{i+1} \subseteq A_i$ . Note we intersect with  $[0, 1]$  because in the beginning,  $c - 1$  to  $c + 1$  may contain stuff outside of the interval. Actually, if  $c = 0$  or  $c = 1$  all sets would contain stuff outside of  $[0, 1]$ . Because  $P([a, b]) = b - a$ ,  $P(A_i) = c + \frac{1}{i} - c + \frac{1}{i} = \frac{2}{i}$ . So note that:

$$\bigcap_{i=1}^{\infty} A_i = \{c\}.$$

So,

$$P(\{c\}) = P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \rightarrow \infty} P(A_i) = \lim_{i \rightarrow \infty} \frac{2}{i} = 0.$$

- So what are some common fields. Suppose  $\Omega = \mathfrak{R}$ . What is the smallest field associated with  $\Omega$ ? Answer:

$$\mathfrak{A} = \{\emptyset, \mathfrak{R}\}.$$

- Lemma 1: The intersection of (possibly uncountable many)  $\sigma$ -algebras is also a  $\sigma$ -algebra in  $\Omega$ .
- Lemma 2: Let  $\mathfrak{E}$  be an arbitrary set of subsets of  $\Omega$  (not necessarily an algebra). For example,

$$\mathfrak{E} = \{(a, b], a \leq b, a, b \in \mathfrak{R}\}.$$

$\mathfrak{E}$  would not be a  $\sigma$ -algebra because it does not contain the complement. Now suppose we have a bunch of  $\sigma$ -algebras in  $\Omega$  that all contain  $\mathfrak{E}$ . So:

$$\mathfrak{E} \subseteq \mathfrak{A}_1, \mathfrak{E} \subseteq \mathfrak{A}_2, \dots$$

Let:

$$\sigma(\mathfrak{E}) = \bigcap_{i=1}^{\infty} \mathfrak{A}_i.$$

So  $\sigma(\mathfrak{E})$  is the intersection of  $\sigma$ -algebras so it is also a  $\sigma$ -algebra. Clearly it is also the SMALLEST  $\sigma$ -algebra that contains  $\mathfrak{E}$ .

- Remark: Denote  $\mathfrak{F}$  as one of the  $\sigma$ -algebras that contains  $\mathfrak{E}$ . Then of course  $\sigma(\mathfrak{E}) \subseteq \mathfrak{F}$ .
- **Definition:** Intervals in  $\mathfrak{R}^n$ . Denote  $\mathfrak{Z}_1$ , the set of all intervals in  $R^1$  which are open on the left, closed on the right as we defined in our example above. Denote  $\mathfrak{Z}_n$ , the set of all intervals in  $R^n$  which are open on the left, closed on the right.
- **Definition:** Borel Sets in  $\mathfrak{R}^n$ . The smallest  $\sigma$ -algebra containing  $\mathfrak{Z}_n$  is  $\sigma(\mathfrak{Z}_n)$  and it is called the  $\sigma$ -algebra of Borel Sets in  $\mathfrak{R}^n$  and we will denote it as  $\mathfrak{B}^n$ . So if

$$\mathfrak{Z}_1 = \{(a, b] : a, b \in \mathfrak{R}\},$$

then:

$$\begin{aligned} (a, b] &\in \sigma(\mathfrak{Z}_1), \\ (-\infty, a] \cup (b, \infty) &\in \sigma(\mathfrak{Z}_1), \end{aligned}$$

as are intersections and singletons. Note that the borel set is smaller than the powerset but is huge otherwise. We don't use the powerset because it is difficult to define the probability measure on the powerset.

## 4 Lecture 4: September 14, 2004

- Proof example. Suppose we would like to prove that:

$$\lim_{n \rightarrow \infty} \frac{1}{n} = 0.$$

First define Limit: Let  $(a_n)$  be a sequence. Then  $a$  is called the limit of  $(a_n)$  if  $\forall \epsilon > 0$   
 $\exists$  an index  $N_\epsilon$  s.t.:

$$|a_n - a| \leq \epsilon \forall n \geq N_\epsilon.$$

Then

$$\lim_{n \rightarrow \infty} (a_n) = a.$$

So let  $a_n = \frac{1}{n}$ . Let  $[x]$  be the smallest integer that is greater than  $x$ . Thus if  $x = 1.7$ ,  
 $[x] = 2$ . Suppose  $\epsilon > 0$ . Need to show an index exists such that,

$$|a_n - a| = |a_n - 0| = |a_n| = \left| \frac{1}{n} \right| = \frac{1}{n} \leq \epsilon \forall n \geq N_\epsilon.$$

Let  $N_\epsilon = \left[ \frac{1}{\epsilon} \right]$ . So,

$$\frac{1}{\epsilon} \leq \left[ \frac{1}{\epsilon} \right] = N_\epsilon.$$

So,

$$\epsilon \geq \frac{1}{N_\epsilon}.$$

Thus, given an  $\epsilon > 0$ , if we select our index  $N_\epsilon = \left[ \frac{1}{\epsilon} \right]$  then it is clear that:

$$n \geq N_\epsilon \implies \frac{1}{n} \leq \frac{1}{N_\epsilon} \leq \epsilon.$$

So,

$$\frac{1}{n} \leq \epsilon \forall n \geq N_\epsilon.$$

As required. QED.

### 4.1 Probability

- Let  $\mathfrak{Z} = \{(a, b], a \leq b, a, b \in \mathfrak{R}\}$ . And let  $\sigma(\mathfrak{Z})$  be the smallest  $\sigma$ -algebra containing  $\mathfrak{Z}$ . We know that at least one  $\sigma$ -algebra exists : The power set. Call  $\sigma(\mathfrak{Z}) = \mathfrak{B}$ : the  $\sigma$ -algebra of borel sets which is smaller than the power set. Suppose  $(\mathfrak{R}, \mathfrak{B}, P)$  is a probability space and  $P^0$  is another probability measure. And suppose:

$$P((-\infty, a]) = P^0((-\infty, a]), \quad a \in \mathfrak{R}.$$

This implies  $P(B) = P^0(B)$  for  $B \in \mathfrak{B}$ .

- See notes for Lebesgue-Borel Measure. Note the lebesgue integral will not in general equal the Reimann integral in all cases.

## 4.2 Random Variables

- Consider again the 2-coin flip example.

$$\Omega = \{TT, TH, HT, HH\}.$$

$$\Omega^+ = \{0, 1, 2\}.$$

$$X : \Omega \mapsto \Omega^+.$$

Denote:

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = TT \\ 1 & \text{if } \omega = TH, \omega = HT \\ 2 & \text{if } \omega = HH \end{cases}$$

So in words,  $X$  is the number of heads we see in our 2 flips. Next define:

$$A = \{TT, HT\}.$$

Define the image of  $A$  as:

$$X(A) = \{X(\omega) \in \Omega^+ : \omega \in A\} = \{0, 1\} = A^+.$$

See G-4.1 in notes. Next define the inverse image of  $A$  as:

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A^+\} = \{TT, TH, HT\}.$$

Now suppose we have probabilities for  $\omega$  so  $(\Omega, P) \mapsto (\Omega^+, P^+)$ . Thus,

$$P^+(A^+) = P^+(\{0, 1\}) = P(X^{-1}(\{0, 1\})).$$

This is all well defined because the inverse image is in the original set (the power set):

$$\mathfrak{A} = \mathcal{P}(\Omega) \equiv \text{Power Set}.$$

But now suppose we have  $\Omega$ ,  $P$ , and:

$$\mathfrak{A} = \{\emptyset, \Omega, \{TT\}, \{TH, HT, HH\}\}.$$

And  $\Omega^+$ ,  $P^+$ ,  $\mathfrak{A}^+ = \mathcal{P}(\Omega^+)$ . Now,

$$X^{-1}(A^+) = X^{-1}(\{0, 1\}) = \{TT, HT, TH\} \notin \mathfrak{A}.$$

So in order for the mapping to make sense, we cannot have a coarse original space and a detailed new space. The Inverse image must be in the original set. If it is, we call the mapping measurable.

- **Definition:** Let  $(\Omega, \mathfrak{A})$  and  $(\Omega^+, \mathfrak{A}^+)$  be two measurable spaces. Let  $X : \Omega \mapsto \Omega^+$  be function. Then  $X$  is said to be  $(\mathfrak{A}, \mathfrak{A}^+)$ -measurable if:

$$X^{-1}(A^+) \in \mathfrak{A} \forall A^+ \in \mathfrak{A}^+.$$

- Suppose  $\Omega = \mathfrak{R}, \mathfrak{A} = \mathfrak{B}, \Omega^+ = \mathfrak{R}, \mathfrak{A}^+ = \mathfrak{B}$  and  $X : \mathfrak{R} \mapsto \mathfrak{R}$ . Then if  $X^{-1}((-\infty, a]) \in \mathfrak{B}$  then  $X$  is Borel-Borel Measurable.

- **Theorem:** Let  $(\Omega, \mathfrak{A})$  be some measurable space. A function  $X : \Omega \mapsto \mathfrak{R}^n$  must have the form:

$$X(\omega) = (X_1(\omega), \dots, X_n(\omega))$$

with  $X_i : \Omega \mapsto \mathfrak{R}$ . Then  $X$  is  $(\mathfrak{A}, \mathfrak{B}^n)$ -measurable iff each component  $X_i$  is  $(\mathfrak{A}, \mathfrak{B})$ -measurable.

- **Definition:** A function  $X : \mathfrak{R}^n \mapsto \mathfrak{R}^m$  is called measurable if it is  $(\mathfrak{B}^n, \mathfrak{B}^m)$ -measurable. Such a function  $X$  is often called a Borel-Function.

- If  $X : \mathfrak{R}^n \mapsto \mathfrak{R}^m$  is continuous, then it is measurable.

- **Theorem:** If two functions,  $X$  and  $Y$  are measurable, then their composite function,  $Z$ , is also measurable. This implies that sums, products, and maxima of measurable functions are measurable.

- **Definition:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space and  $(\Omega^+, \mathfrak{A}^+)$  be a measurable space. Let  $X : \Omega \mapsto \Omega^+$  be a  $(\mathfrak{A}, \mathfrak{A}^+)$ -measurable function from  $\Omega$  to  $\Omega^+$ . We then say that  $X$  is a random variable that takes its values in  $\Omega^+$ .

If  $\Omega^+ \subseteq \mathfrak{R}$ ,  $X$  is a real valued random variable.

If  $\Omega^+ \subseteq \overline{\mathfrak{R}}$ ,  $X$  is an extended real valued random variable.

If  $\Omega^+ \subseteq \mathfrak{R}^n$ ,  $X$  is an  $n$ -dimensional real valued random vector.

If  $\Omega^+ \subseteq \overline{\mathfrak{R}^n}$ ,  $X$  is an  $n$ -dimensional extended real valued random vector.

- Consider the following:  $(\Omega = \mathfrak{R}, \mathfrak{A} = \mathfrak{B}), (\Omega^+ = \mathfrak{R}, \mathfrak{A}^+ = \mathfrak{B})$  and  $X : \mathfrak{R} \mapsto \mathfrak{R}$ . Suppose:

$$X^{-1}((-\infty, a]) \in \mathfrak{B}.$$

Note this means that  $X$  is a random variable. If,

$$X^{-1}(B) \in \underbrace{\mathfrak{B}}_{\mathfrak{A}} \forall B \in \underbrace{\mathfrak{B}}_{\mathfrak{A}^+},$$

then  $X$  is  $(\mathfrak{B}, \mathfrak{B})$ -measurable. We can also get this by the sufficient condition in Corollary 1 of section II.1:  $X$  is  $(\mathfrak{B}, \mathfrak{B})$ -measurable if:

$$X^{-1}((-\infty, a]) \in \mathfrak{B} \forall (-\infty, a].$$

- **Definition:** Induced Probability Measure. Let  $(\Omega, \mathfrak{A}, P)$  be a probability space and let  $(\Omega^+, \mathfrak{A}^+)$  be a measurable space. Let  $X : \Omega \mapsto \Omega^+$  be a  $(\mathfrak{A}, \mathfrak{A}^+)$ -measurable function, i.e. a random variable taking its values in  $\Omega^+$ . Then the function  $P_x : \mathfrak{A}^+ \mapsto [0, 1]$  defined as:

$$P_x(A^+) = P(X^{-1}(A^+)) = P(\{\omega \in \Omega : X(\omega) \in A^+\}) \quad \forall A^+ \in \mathfrak{A}^+,$$

is called the  $X$  **induced probability measure**, or the distribution of the random variable  $X$ , or the law of the random variable  $X$ .

## 5 Lecture 5: September 16, 2004

- Note from last lecture: A random variable is a function which maps the real line into  $\mathfrak{R}^n$ . There should be measurable information sets such that you can carry probabilities back and forth between the spaces. Should have measurability with respect to the two sets.

### 5.1 Distribution Functions

- Let  $X$  be a random variable. Define:

$$F(x) = Pr(X \leq x).$$

- Lemma 1: Let  $P_x$  be a probability measure on  $(\mathfrak{R}, \mathfrak{B})$ . Then  $F : \mathfrak{R} \mapsto [0, 1]$  defined as:

$$F(x) = P_x((-\infty, x]) \quad \forall x \in \mathfrak{R},$$

has the following properties:

- 1)  $F(x)$  is non-decreasing. If  $x_1 \leq x_2$ , then  $F(x_1) \leq F(x_2)$ . This is clear because:

$$F(x_1) = P_x((-\infty, x_1]).$$

$$\begin{aligned} F(x_2) &= P_x((-\infty, x_2]) = P((-\infty, x_1] \cup (x_1, x_2]) = P((-\infty, x_1]) + P((x_1, x_2]) = \\ &= F(x_1) + \underbrace{P((x_1, x_2])}_{\geq 0} \geq F(x_1). \end{aligned}$$

- 2)  $\lim_{x \rightarrow \infty} F(x) = F(\infty) = 1$ ,  $\lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0$ . To show the second part, consider a sequence  $(x_n)$  which converges to negative infinity and consider:

$$\bigcap_{n=1}^{\infty} (-\infty, x_n] = \emptyset.$$

As  $n$  gets larger, we get smaller and smaller sets so the intersection would just be the left hand boundry but since it is not included in the set, we don't get anything at all. So,

$$P(\emptyset) = 0 = \lim_{n \rightarrow \infty} Pr((-\infty, x_n]) = F(x_n).$$

The first part is clear because as  $x \rightarrow \infty$ , we get the entire real line and since probability measures are bounded above by 1, we must get the result.

- 3)  $F$  is right-continuous.  $F(x+) = \lim_{x \searrow 0} F(x+h) = F(x) \quad \forall x \in \mathfrak{R}$ . So the limit of a sequence that converges to  $x$  from the right must be equal to the value of the function itself. We don't necessarily need it be left-continuous, or indeed, a totally continuous function. This means, since  $F(x)$  is non-decreasing, if there is jump in the function, the value of the function at the point must be on top. To

see this, consider:

$$\begin{aligned} F(x + h_n) &= Pr((-\infty, x + h_n]) = Pr((-\infty, x] \cup (x, h_n]) = \\ &= Pr((-\infty, x]) + Pr((x, h_n]) = F(x) + Pr((x, h_n]). \end{aligned}$$

Now let  $A_n = (x, h_n]$ . Since  $h_n \searrow 0 \implies A_{n+1} \subseteq A_n$ . Thus,

$$\bigcap_{n=1}^{\infty} A_n = \emptyset.$$

Again,  $x$  is not in the interval so the intersection is empty. Thus,

$$Pr(A_n) = Pr((x, h_n]) = 0.$$

Thus,

$$F(x + h_n) = F(x) + \underbrace{Pr((x, h_n])}_0 = F(x).$$

- **Definition:** If  $F$  has the properties above, call  $F(x)$  a Distribution Function on the real line.
- **Theorem:** Let  $F : \mathfrak{R} \mapsto [0, 1]$  be a distribution function. Then there exists a unique probability measure,  $P_*$ , on  $(\mathfrak{R}, \mathfrak{B})$  such that:

$$P_*((a, b]) = F(b) - F(a) \geq 0 \text{ for } a, b \in \mathfrak{R} \text{ with } a < b.$$

- Let  $X : (\Omega, \mathfrak{A}, P) \mapsto (\mathfrak{R}, \mathfrak{B}, P_x)$  be a real valued random variable. [[What we mean by this is that  $X : \Omega \mapsto \mathfrak{R}$ ,  $\mathfrak{A}$  is a  $\sigma$ -algebra (information set) in the domain,  $\mathfrak{B}$  is the information set in the range,  $P$  is the probability measure in the domain ( $P : \mathfrak{A} \mapsto [0, 1]$ ), and  $P_x$  is the INDUCED probability in the range ( $P_x : \mathfrak{B} \mapsto [0, 1]$ ).] Then the function  $F : \mathfrak{R} \mapsto [0, 1]$  defined as:

$$\begin{aligned} F(x) &= Pr(X \leq x) = Pr\{X \in (-\infty, x]\} = P(\{\omega \in \Omega : X(\omega) \in (-\infty, x]\}) = \\ &= \underbrace{P(X^{-1}((-\infty, x]))}_{\subseteq \Omega} = P(\{\omega \in \Omega : X(\omega) \in A^+\}) = P_x((-\infty, x]), \end{aligned}$$

is called the CUMULATIVE distribution function of  $X$ . So  $P$  operates on subsets of  $\Omega$  and  $P_x$  operates on subsets of the real line.

- Corollary 1:

$$\begin{aligned} P(a < X \leq b) &= P(\{\omega \in \Omega : a < X(\omega) \leq b\}) = P(X^{-1}((a, b])) = \\ &= P_x((a, b]) = F(b) - F(a) \geq 0. \end{aligned}$$

Also,

$$P(X = b) = P_x(\{b\}).$$

To see this, define:

$$\{b\} = \bigcap_{n=1}^{\infty} (b - \frac{1}{n}, b].$$

Thus,

$$\begin{aligned} P_x(\{b\}) &= P_x\left[\bigcap_{n=1}^{\infty} (b - \frac{1}{n}, b]\right] = \lim_{n \rightarrow \infty} P_x((b - \frac{1}{n}, b]) = \\ &= \lim_{n \rightarrow \infty} (F(b) - F(b - \frac{1}{n})) = F(b) - F(b-). \end{aligned}$$

So we have  $F(b)$  less the distribution function of the point  $b$  coming from the left. Since  $F$  is only right continuous,  $F(b-)$  may not equal  $F(b)$ . If  $F$  is continuous at  $b$ , then  $P(X = b) = 0$ , but if there is a point of discontinuity,  $F(b) - F(b-)$  is equal to the height of the jump. See G-5.1.

- **Theorem:** Decomposition of Distribution Function.  $X$  is a random variable  $\sim N(0, 1)$ .  $Y$  is a random variable such that:  $Y = 0$  if  $X \leq 0$ ,  $Y = X$  if  $X > 0$ . See G-5.2 for the distribution functions of  $X$  and  $Y$ . At the point  $y = 0$ ,  $F(y)$  jumps to 0.5 because,  $P(Y \leq y)$  is 0.5. Thus we can write,

$$F(y) = \frac{1}{2}F_1(y) + \frac{1}{2}F_2(y),$$

where,

$$F_1(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0 \end{cases}$$

and,

$$F_2(y) = \begin{cases} 0 & y < 0 \\ F_{N(0,1)}(y) & y \geq 0 \end{cases}$$

So the theorem says: let  $F : \Re \mapsto [0, 1]$  be the cumulative distribution function of the real valued random variable,  $X$ . Then  $F$  can be uniquely decomposed in the form:

$$F(x) = pF_1(x) + (1 - p)F_2(x), \quad 0 \leq p \leq 1.$$

Where  $F_1$  and  $F_2$  are distribution functions and  $F_1$  is a step function while  $F_2$  is continuous everywhere. Further,  $F_2$  can be decomposed as:

$$F_2(x) = qF_{21}(x) + (1 - q)F_{22}(x), \quad 0 \leq q \leq 1.$$

Where  $F_{21}$  is absolutely continuous, ie,  $F_{21}(x) = \int_{-\infty}^x f(t)dt, f(t) > 0$ . And  $F_{22}$  is continuous singular: which means it is continuous and has derivative = 0 almost everywhere (We cannot write it as an integral over some interval).

- In multiple dimensions, it is easy enough to extend the analysis. Consider:

$$F(x_1, x_2) = Pr(X_1 \leq x_1, X_2 \leq x_2).$$

Or more interestingly,

$$Pr(a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2) = F(b_1, b_2) + F(a_1, a_2) - F(a_1, b_2) - F(a_2, b_1).$$

See graph G-5.3 for why. The idea is, we start with  $F(b_1, b_2)$  and remove  $F(a_1, b_2)$  and  $F(a_2, b_1)$  which is all good but we have removed  $F(a_1, a_2)$  twice! So add it back in.

## 6 Lecture 6: September 21, 2004

### 6.1 Review of Distribution Functions

- Recall the distribution function for a 2x2 space:

$$\begin{aligned} P(\{a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2\}) &= P(\{(x_1, x_2) \in (a_1, b_1] \times (a_2, b_2]\}) = \\ &= P_x((a_1, b_1], (a_2, b_2]) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2). \end{aligned}$$

Where  $F_x(x_1, x_2) = Pr(X_1 \leq x_1, X_2 \leq x_2)$ .

- There is a one-to-one relationship between the probability law for random variables,  $P_x$ , and the distribution function of  $X$ ,  $F_x$ .

### 6.2 Probability Density Function

- **Definition 1** A real valued random variable is called discrete if it only assumes countable many values. Its corresponding distribution function,  $F_x$ , is called discrete.
- **Definition 2** Let  $X$  be a discrete RV with distinct values  $x^1, x^2, \dots$ . Then the function:

$$f_x(x) = \begin{cases} Pr(X = x^i) & x = 1, 2, \dots \\ 0 & \text{if } x \neq x^i \end{cases}$$

$f_x(x)$  is called the discrete probability density function of  $X$  (PDF).

- Example. Let  $x^i = i$ , for  $i = 1, 2, \dots$ . Then,

$$f_x(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x = 1, 2, \dots \\ 0 & \text{else} \end{cases}$$

See G-6.1 for a picture of this “step” PDF.

- **Remark** Cumulative Density Function (CDF).

$$F_x(x) = Pr(X \leq x) = \sum_{x^i \in (-\infty, x)} Pr(X = x^i) = \sum_{x^i \in (-\infty, x)} f_x(x^i).$$

So the CDF in the discrete case is just the sum of the discrete PDF at all values up to a point  $x$ . Also note:

$$Pr(X = x^i) = f_x(x^i) = P_x(\{x^i\}) = F(x^i) - F(x^i -).$$

So since  $F_x(x)$  is a step function, the height of the step is the probability with which  $X$  takes on a value  $x^i$ .

- More properties of the PDF.

$$0 \leq f_x(x) \leq 1.$$

$$\sum_{x^i \in (-\infty, \infty)} f_x(x^i) = 1.$$

$$Pr(X \in A) = P_x(A) = \sum_{x^i \in A} f_x(x^i).$$

- Consider again our example above and let  $A = \{x = 1, 3, 5, 7, \dots\}$ . Thus,

$$\begin{aligned} P(X \in A) &= \sum_{x=1,3,5,\dots} f_x(x^i) = \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \dots = \\ &= \sum_{z=0}^{\infty} \left(\frac{1}{2}\right)^{2z+1} = \frac{1}{2} \sum_{z=0}^{\infty} \left[\left(\frac{1}{2}\right)^2\right]^z = \\ &= \frac{1}{2} * \frac{1}{1 - 1/4} = \frac{1}{2} * \frac{4}{3} = \frac{2}{3}. \end{aligned}$$

- For the discrete case, the density can be interpreted as the probability of the event happening so  $f_x(x) \in [0, 1]$ , but for continuous random variables, this is not the case.
- **Definition 3** A real valued random variable is called a continuous RV if there exists  $f_x(x) \geq 0$  such that the CDF can be expressed as:

$$F_x(x) = \int_{-\infty}^x f_x(t)dt, \quad x \in \mathfrak{R}.$$

If  $X$  is continuous,  $F_x$  is called “Absolutely” Continuous. Then  $f_x(t)$  is the probability density function of  $X$ . See G-6.2

- **Remark 2** Suppose  $X$  is a continuous random variable. Then:

- $F_x$  is continuous everywhere. So the probability of  $X$  at any point, say  $x^0$ , is zero.

$$Pr(X = x^0) = P_x(\{x^0\}) = F(x^0) - F(x^0-) = 0.$$

- Properties:  $0 \leq f_x(x) < \infty$ ,  $\int_{-\infty}^{\infty} f_x(x)dx = 1$ ,

$$Pr(X \in A) = P_x(A) = \int_A f_x(x)dx.$$

And note that:

$$f_x(x) = \frac{\partial F_x(x)}{\partial x},$$

at every continuous point of  $f_x(x)$ .

- Suppose

$$F_x(x) = \int_{-\infty}^x g_x(t)dt,$$

then  $f_x = g_x$ , almost everywhere. See G-6.3 for a picture of this. The PDF can be different at a given point and have this equality still hold since the probability of any given point is 0. Thus the PDF is NOT unique in the continuous case. The inclusion/exclusion of end points also will not matter for the same reason.

- Consider another example:

$$f_x(x) = \begin{cases} cx^2 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

Then,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^0 f(x)dx + \int_0^1 f(x)dx + \int_1^{\infty} f(x)dx = \\ &\int_0^1 f(x)dx = \int_0^1 cx^2 dx = \frac{1}{3}cx^3 \Big|_0^1 = \frac{1}{3}c. \end{aligned}$$

So  $c = 3$  for this to be a PDF.

- Another example. See G-6.4.

$$f_x(x) = \begin{cases} cx^{-2} & 1 < x < \infty \\ 0 & \text{else} \end{cases}$$

Thus,

$$1 = \int_1^{\infty} cx^{-2} dx = -cx^{-1} \Big|_1^{\infty} = c.$$

So  $c = 1$  for this to be a PDF. We can use this example to give a little look into the next topic, expectation. Given the PDF in this example, define:

$$E[x] = \int_{-\infty}^{\infty} xf(x)dx = \int_1^{\infty} x \cdot x^{-2} dx = \int_1^{\infty} \frac{1}{x} dx = \ln(x) \Big|_1^{\infty} = \infty.$$

So the expected value of  $x$  (the first moment of  $x$ ) is infinity.

- Another example.

$$f_x(x) = \begin{cases} \frac{1}{2}x^{-2} & -\infty < x < -1, 1 < x < \infty, \\ 0 & \text{else} \end{cases}$$

Note this has to be a PDF because we have doubled the domain, but halved the density.

Thus,

$$E[x] = \int_{-\infty}^{\infty} \frac{1}{2}x \cdot x^{-2} dx = \int_{-\infty}^{\infty} \frac{1}{2}x^{-1} dx = \infty - \infty = DNE.$$

SO NOT ALL RANDOM VARIABLES HAVE AN EXPECTED VALUE!

### 6.3 Mathematical Expectation

- **Definition 1a:** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space and let  $X : \Omega \mapsto \mathfrak{R}$  be a scalar RV with a distribution  $P_x$  and let  $u : \mathfrak{R} \mapsto \mathfrak{R}$  be a  $(\mathfrak{B}, \mathfrak{B})$ -measurable function.

- Suppose  $X$  is a discrete RV with PDF,  $f_x(x)$ , then the expected value (mathematical expectation) of  $u(x)$  is defined as:

$$E[u(x)] = \sum_{x^i} u(x^i) f_x(x^i),$$

where  $P_x\{x^1, x^2, \dots\} = 1$ . Note that if  $u(x) = x$ ,

$$E[x] = \sum_{x^i} x^i f_x(x^i),$$

If  $u(x) = x^2$ ,

$$E[u(x)] = E[x^2] = \sum_{x^i} (x^i)^2 f_x(x^i),$$

or the second moment of  $x$  (the variance if  $E[x] = 0$ .)

- Now suppose  $X$  is a continuous RV. Then:

$$E[u(x)] = \int_{-\infty}^{\infty} u(x) f_x(x) dx.$$

Or for  $u(x) = x$ ,

$$E[x] = \int_{-\infty}^{\infty} x f_x(x) dx.$$

- Consider a discrete RV with PDF:

$$f_x(x) = \begin{cases} \frac{1}{2} & x = -1, x = 1 \\ 0 & \text{else} \end{cases}$$

Then  $E[x] = 0$  but 0 is not a possible outcome.

- Finally, we have a more general formula for expectation when we don't necessarily have to have a continuous or a discrete RV:

$$E[x] = \int_0^{\infty} [1 - F_x(x)] dx - \int_{-\infty}^0 F_x(x) dx.$$

## 7 Lecture 7: September 23, 2004

### 7.1 Reimann and Lebesgue Integrals

- Consider a continuous function  $g(x)$  on the interval  $[a, b]$ . Define a lower summation:

$$L_{\mathfrak{Z}_n} = \sum_{i=1}^n [\inf_{x_{i-1} < x \leq x_i} g(x)](x_i - x_{i-1}).$$

Where  $\mathfrak{Z}_n = x_0 < x_1 < x_2 < \dots < x_n$ . And also an upper summation:

$$U_{\mathfrak{Z}_n} = \sum_{i=1}^n [\sup_{x_{i-1} < x \leq x_i} g(x)](x_i - x_{i-1}).$$

Note we use infimum and supremum instead of minimum and maximum because the function may not be defined at the min or max. Hence use *inf* and *sup* to denote the limit approaching the min or max. Then we have, if:

$$\lim_{n \rightarrow \infty} L_{\mathfrak{Z}_n} = \lim_{n \rightarrow \infty} U_{\mathfrak{Z}_n},$$

then, these are both equal to:

$$(R) \int_a^b g(x) dx.$$

The Reimann Integral. Note that as  $n$  went to infinity, we just summed over more rectangles.

- Now rewrite the problem and let  $A_i = (x_{i-1}, x_i]$ . Then the Lebesgue measure is  $(x_i - x_{i-1})$ . And we can write from above:
- Consider a continuous function  $g(x)$  on the interval  $[a, b]$ . Define a lower summation:

$$L_{\mathfrak{Z}_n} = \sum_{i=1}^n [\inf_{x \in A_i} g(x)] \lambda(A_i).$$

$$U_{\mathfrak{Z}_n} = \sum_{i=1}^n [\sup_{x \in A_i} g(x)] \lambda(A_i).$$

And this is completely equivalent to the Reimann integral. Note that,

$$[a, b] = \bigcup_{i=1}^n A_i, \quad A_i \cap A_j = \emptyset \text{ for } i \neq j.$$

But with the Lebesgue integral, we don't have to define the  $A_i$ 's like we did above. Any disjoint partition will do as long as their union gives us the whole interval. Therefore,

we define the Lebesgue Integral as:

$$L \int_a^b g(x)dx = \sup_Q \sum_{i=1}^n [\inf_{x \in A_i} g(x)] \lambda(A_i),$$

where  $Q$  is ALL finite partitions of  $[a, b]$ . Since we are taking the infimum on the inside, we know we are never exceeding the area under  $g(x)$  for any partition, and when we take the supremum over all finite partitions, we should get the actual integral (Lebesgue). This is sometimes denoted as:

$$\int g d\lambda.$$

So in general,

$$\int g d\mu = \int g(x) d\mu(x) = \int g(x) \mu(dx)$$

are all just notation changes where  $\mu$  can be any measure (probability included as in expected value).

- Example. Consider a function  $g(x) = \mathbf{1}_A(x)$ , an indicator variable that equals 1 if  $x \in A$  and equals 0 if  $x \in A^c$  where,

$A$  = the set of irrationals on  $[0, 1]$ .

$A^c$  = the set of rationals on  $[0, 1]$ .

So this looks like a line at  $g(x) = 1$  on the interval  $[0, 1]$  with countably many holes in it (1 for each rational number). Note that infinite sets like all the counting numbers, or all rational numbers, both countable sets, have zero measure. Thus  $\lambda(A^c) = 0$ . Since the rationals are countable, we can take the infinite union of all the rationals in the interval, but each individual rational has length 0 so the set itself has length 0. For the irrationals, these are uncountable so the lebesgue measure is equal to 1, the length of the entire interval. I think that makes sense. So,

$$\lambda(A) = 1,$$

$$\lambda(A^c) = 0.$$

Also note that  $\inf_{x \in A} g(x) = 1$  and  $\inf_{x \in A^c} g(x) = 0$ . Thus, applying the definition of the Lebesgue integral:

$$L \int_0^1 g(x)dx = \sum_{i=1}^2 [\inf_{x \in A_i} g(x)] \lambda(A_i) = 1 * \lambda(A) + 0 * \lambda(A^c) = 1.$$

Note that the Riemann integral would have not been able to solve this.

## 7.2 Returning to Mathematical Expectation

- **Remark** The expected value of a random variable  $X$  is defined as the integral of  $X$  with respect to the measure  $P$  (given it exists):

$$E[X] = \int X dP = \int X(\omega) dP(\omega) = \int u(x) dP_x(x).$$

In the case where  $X : \Omega \mapsto \Re$  is a scalar random variable with distribution function  $F_x$ , it can be shown:

$$E[X] = \int_0^\infty [1 - F_x(x)] dx - \int_{-\infty}^0 F_x(x) dx.$$

- Now consider the 2-dimension case:  $X = (x_1, x_2)$ . Thus,

$$E[u(x_1, x_2)] = \int \int u(x_1, x_2) f(x_1, x_2) dx_1 dx_2.$$

Consider the special case where  $u(x_1, x_2) = x_1$ ,

$$E[x_1] = \int \int x_1 f(x_1, x_2) dx_1 dx_2 = \int x_1 \underbrace{\int f(x_1, x_2) dx_2}_{f_1(x_1)} dx_1 = \int x_1 f_1(x_1) dx_1.$$

Note that  $f(x_1, x_2)$  is the joint density of  $x_1$  and  $x_2$ , while  $f_1(x_1)$  is the marginal density of  $x_1$ .

- Consider a random variable that is neither discrete nor continuous. We have seen we can write its distribution function as:

$$F_x(x) = pF_1(x) + (1 - p)F_2(x), \quad 0 < p < 1.$$

Where  $F_1$  is a jump function and  $F_2$  is continuous:

$$F_1(x) = \sum_{x^i < x} f_1(x^i),$$

$$F_2(x) = \int_{-\infty}^x f_2(t) dt.$$

Then it is easy to see:

$$E[u(x)] = p \sum_{x^i} u(x^i) f_1(x^i) + (1 - p) \int_{-\infty}^{\infty} u(x) f_2(x) dx.$$

- Example. Suppose  $Y$  is continuous with CDF,  $F_y(y)$ , and PDF,  $f_y(y)$ . Define  $X$  as:

$$X = \begin{cases} c & Y \leq c \\ Y & Y \geq c \end{cases}$$

Thus the CDF of  $X$  is:

$$F_x(x) = Pr(X \leq x) = \begin{cases} 0 & x < c \\ F_y(x) & x \geq c \end{cases}$$

We can split  $F_x(x)$  into two parts:

$$F_x(x) = pF_1(x) + (1 - p)F_2(x),$$

with,

$$p = Pr(Y \leq c) = F_y(c).$$

$$F_1(x) = \begin{cases} 0 & x < c \\ 1 & x \geq c \end{cases}$$

$$F_2(x) = \begin{cases} 0 & x < c \\ \frac{F_y(x) - F_y(c)}{1 - F_y(c)} & x \geq c \end{cases}$$

If you mess with the numbers, you see that for  $x < c$ ,  $F_x(x) = 0$ , but if  $x \geq c$ , then we have something like:

$$F_y(c) + (1 - F_y(c)) \frac{F_y(x) - F_y(c)}{1 - F_y(c)} = F_y(x).$$

So this seems to work. The corresponding PDFs are:

$$f_1(x) = \begin{cases} 1 & x = c \\ 0 & x \neq c \end{cases}$$

$$f_2(x) = \begin{cases} 0 & x < c \\ \frac{f_y(x)}{1 - F_y(c)} & x \geq c \end{cases}$$

The last equation coming from differentiating  $F_2(x)$  with respect to  $x$ . Hence applying the formula above for  $u(x) = x$ ,

$$E[u(x)] = p \sum_{x^i} u(x^i) f_1(x^i) + (1 - p) \int_{-\infty}^{\infty} u(x) f_2(x) dx.$$

$$E[x] = p \sum_{x^i} x^i f_1(x^i) + (1-p) \int_{-\infty}^{\infty} x f_2(x) dx.$$

$$E[x] = F_y(c) * c + (1 - F_y(c)) \int_c^{\infty} x \frac{f_y(x)}{1 - F_y(c)} dx.$$

$$E[x] = Pr(Y \leq c) * c + \int_c^{\infty} x f_y(x) dx.$$

- **Theorem:** Suppose  $X$  and  $Y$  are real valued random variables on a probability space  $(\Omega, \mathfrak{A}, P)$  and let  $a, b \in \mathfrak{R}$ . Then:

- 1)  $E[a] = \int a f(x) dx = a \int f(x) dx = a.$
- 2)  $E[ax] = \int ax f(x) dx = a \int x f(x) dx = aE[x].$
- 3)

$$\begin{aligned} E[ax + by] &= \int \int (ax + by) f(x, y) dx dy \\ &= \int \int ax f(x, y) dx dy + \int \int by f(x, y) dx dy \\ &= a \int x \int f(x, y) dy dx + b \int y \int f(x, y) dx dy \\ &= a \int x f_x(x) dx + b \int y f_y(y) dy \\ &= aE(x) + bE(y). \end{aligned}$$

- **Remark** In general  $E[u(x, y)] \neq u(E[x], E[y])$ . Consider,

$$f_x(x) = \begin{cases} \frac{1}{2} & x = -1, x = 1 \\ 0 & \text{else} \end{cases}$$

Then  $E[X] = 0$ . Now let  $Y = X$ .  $E[XY] = E[X^2] = 1$ . However  $E[X] * E[Y] = 0 * 0 = 0$ .

- **Theorem:** Let  $X, Y, X_1, X_2, \dots$  represent real valued RVs on a probability space  $(\Omega, \mathfrak{A}, P)$ . Then,

- 1) If  $X = 0$  almost everywhere, then  $E[X] = 0$ .
- 2) If  $X$  is non-negative and  $P(\{\omega : X(\omega) > 0\}) > 0$ , then  $E[X] > 0$ .
- 3) If  $E[|X|] < \infty$ , then  $|E[X]| \leq E[|X|]$ .
- 4) If  $E[|X|] < \infty$ , then  $|X| < \infty$  almost everywhere.
- 5) (Monotonicity) If  $E[|X|] < \infty$  and  $E[|Y|] < \infty$  and  $X \leq Y$  almost everywhere, then  $E[X] \leq E[Y]$ .

- 6) If  $E[|X|] < \infty$  and  $E[|Y|] < \infty$  and  $X = Y$  almost everywhere, then  $E[X] = E[Y]$ .
- 7) (Monotone Convergence) If  $0 \leq X_n \uparrow X$  almost everywhere, then  $E[X_n] \uparrow E[X]$ .
- 8) (Lebesgue Dominated Convergence) If  $|X_n| \leq Y$  almost everywhere ( $Y$  dominates  $X$ ) with  $E[|Y|] < \infty$  and if  $X_n \rightarrow X$  almost everywhere, then  $E[|X|] < \infty$ ,  $E[|X_n|] < \infty$ , and  $E[X_n] \rightarrow E[X]$ .

### 7.3 Mean, Variance and Other Characteristics

- **Definition:** Let  $X$  be a random variable. Then assuming the existence of the expectations defined below:

- 1)  $E[X^r]$  is called the  $r^{\text{th}}$  population moment of  $X$ ,  $r \geq 0$ , usually denoted by  $\mu'_r$ .
- 2) For  $r = 1$ ,  $E[X]$  is called the population mean of  $X$ , denoted by  $\mu_x$  or  $\mu'_1$ . Note the first moment, or the population mean, is not equal to the sample mean of  $X$ .

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \text{Sample Mean.}$$

$$E[X] = \int x f(x) dx \equiv \text{Population Mean.}$$

$E[X]$  is a number, the expected value, while the sample mean is itself a random variable. In fact,

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \mu.$$

So the expected value of the sample mean is the population mean.

- 3)  $E[(X - a)^r]$  is the  $r^{\text{th}}$  central moment around  $a$ . Often we set  $a = E[X]$ .
- 4)  $E[(X - \mu_x)^r]$  is the  $r^{\text{th}}$  central moment about  $\mu_x$ . Often denoted  $\mu_r$ .
- 5)  $E[(X - \mu_x)^2]$  is called the variance of  $X$ , usually denoted  $\sigma_x^2$ . Furthermore,  $\sigma_x = \sqrt{\sigma_x^2}$ , the standard deviation of  $X$ . Note also that  $\mu_2 = \sigma_x^2$ .
- 6)  $E[(X - \mu_x)^3]/\sigma_x^3$  is called the skewness of  $X$ . Note Standardization.
- 7)  $E[(X - \mu_x)^4]/\sigma_x^4 - 3$  is called the excess kurtosis of  $X$ . Note Standardization.

## 8 Lecture 8: September 28, 2004

### 8.1 More on Expectation

- Let  $Z = aX$ . Then  $E[Z] = aE[X]$ . Then,

$$\sigma_Z^2 = E[Z - \mu_z]^2 = E[aX - a\mu_x]^2 = a^2 E[X - \mu_x]^2 = a^2 \sigma_x^2.$$

- Kurtosis puts much more weight on the tails but is another measure of tail heaviness. See G-8.1. We often define kurtosis by subtracting 3 from the standardized fourth moment which gives us the excess kurtosis over the normal distribution.
- Proposition 1: Suppose  $X$  is a random variable. Given  $E[|X|^r] < \infty$ , then:

$$E[|X|^s] < \infty \quad \forall 0 < s \leq r.$$

So if a higher moment exists, then all moments of lesser order also exist. Note this is only for ABSOLUTE MOMENTS!

- Proposition 2:  $X$  is a RV with finite variance, then:

$$\begin{aligned}\sigma_x^2 &= E[X - E[X]]^2 \\ &= E[X^2] - 2E[X * E[X]] + E[E[X]^2] \\ &= E[X^2] - 2E[X] * E[X] + [E[X]]^2 \\ &= E[X^2] - 2[E[X]]^2 + [E[X]]^2 \\ &= E[X^2] - [E[X]]^2 \\ &= E[X^2] - \mu_x^2\end{aligned}$$

- **Definition:** Quantile.  $X$  is a RV with CDF,  $F_x(x)$ . Then:

- (1) The  $q^{th}$  quantile of  $X$  is denoted by  $\xi_q$ , with  $0 \leq q \leq 1$ , and is defined as the smallest number,  $\xi$ , such that:

$$F_x(\xi) = Pr(X \leq \xi) \geq q.$$

Or,

$$\xi_q = \inf \{ \xi : F_x(\xi) \geq q \}.$$

See G-8.2 for a depiction. The minimum (*inf*) comes in there because the quantile around a flat area of the CDF will simply be the smallest  $X$  such that the quantile is  $q$ .

- (2) For  $q = 0.5 \implies \xi_q \equiv$  The Median.

- Note also that the Inter-quartile range is often referenced in the literature and is equal to  $\xi_{0.75} - \xi_{0.25}$ .

- **Definition:** Mode. The point  $x$  such that  $f_x(x)$  attains a maximum.
- **Definition:** Suppose  $X$  is a RV. The moment generating function (MGF) of  $X$  is:

$$M(t) = E[e^{tX}] \quad \forall t \in \mathfrak{R} \text{ for which } M(t) \text{ is finite.}$$

- **Theorem:** Suppose  $M(t)$  exists in the interval around zero,  $[-h, h]$ , with  $h > 0$ . Then:

– (1)

$$M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[x^k].$$

– (2) The derivatives of  $M^k(t)$  of all orders  $k$  exist at  $t = 0$  and :

$$M^k(t)|_{t=0} = E[x^k].$$

– (3) The MGF is unique and completely determines the distribution of  $X$ . If  $X$  and  $Y$  have the same MGF, then they have the same probability law as well.

- Therefore we have:

$$M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[x^k].$$

Evaluated at  $t = 0$ , we have:

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{\partial^k M(t)}{\partial t^k} \Big|_{t=0} = \sum_{k=0}^{\infty} \frac{M^k(t)}{k!} t^k \Big|_{t=0}.$$

Expanding:

$$M^1(t) * t + \frac{1}{2 * 1} M^2(t) * t^2 + \frac{1}{3 * 2 * 1} M^3(t) * t^3 + \frac{1}{4 * 3 * 2 * 1} M^4(t) * t^4 + \dots$$

But this is just a Taylor series expansion of the MGF around a point. Here that point is  $t = 0$ , and we also call this a Maclaurin series.

- To show part (2) of the above theorem, consider:

$$M(t) = E[e^{tX}] = \int e^{tX} f(X) dX.$$

So,

$$\begin{aligned} \frac{\partial M(t)}{\partial t} &= \frac{\partial}{\partial t} \int e^{tX} f(X) dX. \\ &= \int \frac{\partial}{\partial t} e^{tX} f(X) dx. \end{aligned}$$

$$= \int X e^{tX} f(X) dx.$$

Evaluated at  $t = 0$ ,

$$= \int X f(x) dx.$$

$$E[X].$$

- Using the theorem, we can rewrite the variance in terms of the MGF as follows:

$$\sigma_x^2 = E[X^2] - [E[X]]^2.$$

$$\sigma_x^2 = M^2(t)|_{t=0} - [M^1(t)]^2|_{t=0}.$$

$$\sigma_x^2 = \left. \frac{\partial^2 M(t)}{\partial t^2} \right|_{t=0} - \left[ \left. \frac{\partial M(t)}{\partial t} \right|_{t=0} \right]^2.$$

- **Definition:** Let  $X$  be a RV. The characteristic function of  $X$  is defined as:

$$\phi(t) = E[e^{itX}].$$

Where  $i = \sqrt{-1}$ . Thus, the characteristic function exists for every distribution for all  $t \in \mathfrak{R}$ . It is unique and completely determines the distribution of a RV.

- Finally, consider an application of MGF to the normal distribution. The PDF of the normal is:

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

It can be shown that the MGF of the normal is:

$$M(t) = e^{t\mu + 0.5\sigma^2 t^2}.$$

Consider the first and second derivatives of  $M(t)$ :

$$\frac{\partial M(t)}{\partial t} = (\mu + t\sigma^2) e^{t\mu + 0.5\sigma^2 t^2}.$$

$$\frac{\partial^2 M(t)}{\partial t^2} = (\mu + t\sigma^2)^2 e^{t\mu + 0.5\sigma^2 t^2} + e^{t\mu + 0.5\sigma^2 t^2} (\sigma^2).$$

Now evaluate both at  $t = 0$ :

$$\frac{\partial M(t)}{\partial t} = (\mu + 0\sigma^2) e^{0\mu + 0.5\sigma^2 0^2} = \mu.$$

$$\frac{\partial^2 M(t)}{\partial t^2} = (\mu + 0\sigma^2)^2 e^{0\mu + 0.5\sigma^2 0^2} + e^{0\mu + 0.5\sigma^2 0^2} (\sigma^2) = \mu^2 + \sigma^2.$$

So applying the formulas from above:

$$E[X] = M^1(t)|_{t=0} = \mu.$$

$$\sigma_x^2 = M^2(t)|_{t=0} - [M^1(t)]^2|_{t=0} = \mu^2 + \sigma^2 - \mu^2 = \sigma^2.$$

## 9 Lecture 9 - September 30, 2004

- Note that we can rewrite the characteristic polynomial formula from last time as:

$$\phi(X) = E[e^{itX}] = E[\cos(tX) + i\sin(tX)].$$

### 9.1 Chebyshev's Inequality

- **Theorem** Suppose  $X$  is a random variable and let  $u : \mathfrak{R} \mapsto [0, \infty)$  be a non-negative function. Then for all  $\epsilon > 0$ ,

$$Pr[u(X) \geq \epsilon] \leq \frac{E[u(X)]}{\epsilon}.$$

Or it is sometimes written:

$$Pr[|X - \mu_x| \geq k\sigma_x] \leq \frac{1}{k^2}.$$

But note the the second formulation follows from the first. Square both sides of LHS:

$$Pr[|X - \mu_x|^2 \geq k^2\sigma_x^2] \leq \frac{E[(X - \mu_x)^2]}{k^2\sigma_x^2} = \frac{\sigma_x^2}{k^2\sigma_x^2} = \frac{1}{k^2}.$$

Where the first  $\leq$  follows from the first formulation of Chebby-Chesnes.

- So what this says, is that given  $\mu_x$  and say an interval of one standard deviation on each side, then the probability that  $X$  falls outside this interval is  $\frac{1}{2}$ , 2 standard deviations:  $\frac{1}{4}$ , 3 standard deviations:  $\frac{1}{9}$ , etc.

- To prove the first formulation consider the following:

$$\begin{aligned}
E[u(x)] &= \int_{-\infty}^{\infty} u(x)f(x)dx \\
&= \int_A u(x)f(x)dx + \int_{A^c} u(x)f(x)dx \\
&\quad \text{Where } A = \{x : u(x) \geq c\} . \\
&= \int_A u(x)f(x)dx + \underbrace{\int_{A^c} u(x)f(x)dx}_{\geq 0} \\
&\quad \text{because } u(x) \text{ is a non-negative function .} \\
&\geq \int_A u(x)f(x)dx \\
&\geq \int_A cf(x)dx \\
&= c \int_A f(x)dx \\
&= c * Pr(X \in A) \\
&= c * Pr(u(x) \geq c) \\
E[u(x)] &\geq c * Pr(u(x) \geq c) \\
Pr(u(x) \geq c) &\leq \frac{E[u(x)]}{c}
\end{aligned}$$

- So consider an application of this inequality. Suppose we consider the sample mean of a random variable and ask if it is a consistent estimate of the population mean. Note that:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim (\mu, \sigma^2/n).$$

So we ask the question, given any  $\epsilon > 0$ , is it the case that:

$$Pr(|\bar{X}_n - \mu| \leq \epsilon) \rightarrow 1 \text{ as } n \rightarrow \infty?$$

Equivalently,

$$Pr(|\bar{X}_n - \mu| \geq \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty?$$

So consider this second formulation:

$$Pr(|\bar{X}_n - \mu| \geq \epsilon) = Pr(|\bar{X}_n - \mu|^2 \geq \epsilon^2).$$

Apply chebychev:

$$Pr(|\bar{X}_n - \mu|^2 \geq \epsilon^2) \leq \frac{E[\bar{X}_n - \mu]^2}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Now, recall that for two sequences such that  $a_n < b_n$  for all  $n$ , we know:

$$\lim_{n \rightarrow \infty} a_n \leq \lim_{n \rightarrow \infty} b_n.$$

[[Note it's only  $\leq$  and not  $<$  because of the case of  $a_n = 0$ ,  $b_n = \frac{1}{n}$ .]] Thus, since probabilities are always non-negative:

$$0 \leq \Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Take the limit as  $n$  goes to infinity:

$$0 \leq \lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq 0.$$

Thus,

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$

So the sample mean is a consistent estimator of the population mean.

- **Theorem** Other Useful Inequalities. Let  $X$  and  $Y$  be RVs. For  $p \geq 1$ , define:

$$\|X\|_p = (E[|X|^p])^{1/p}.$$

This is called the  $L_p$  norm of  $X$ .

- Jensen's Inequality. For a convex function  $g$  on the real line:

$$g(E[X]) \leq E[g(x)].$$

This is simply seen from the simple example where  $X$  equals 0 with probability 0.5 and  $X$  equals 1 with probability 0.5. Then  $E(X) = 0.5$ . For a convex  $g$ , it is easy to see that  $E(g(X)) = 1/2 > g(E(X))$ . G-9.1.

- Lyapunov's Inequality. For  $0 < s \leq r$ ,

$$\|X\|_s \leq \|X\|_r.$$

Or,

$$(E[X]^s)^{1/s} \leq (E[X]^r)^{1/r}$$

This implies that we can put bounds on the norm of a random variable;

$$E[X]^s \leq (E[X]^r)^{s/r}.$$

- Holder's Inequality. For  $\frac{1}{p} + \frac{1}{q} = 1$ ,

$$E[|XY|] = \|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

Or,

$$E[|XY|] \leq (E[X]^p)^{1/p} (E[Y]^q)^{1/q}$$

And for  $p = q = 2$ ,

$$E[|XY|] \leq (E[X]^2)^{1/2} (E[Y]^2)^{1/2}.$$

This last expression is known as the Schwartz Inequality.

– Minkowski's Inequality. This is just a version of the triangle inequality:

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

– Consider the following example of the central limit theorem. Suppose  $X_1 \dots X_n$  are all random variables with unknown distribution and they are completely uncorrelated. Further, suppose:

$$E[X_i^4] \leq K < \infty.$$

So the 4<sup>th</sup> moment exists and is finite. Thus,

$$E[|X_i|] \leq (E[|X_i^4|])^{1/4} \leq K^{1/4} < \infty.$$

Where the first inequality is by Lyapunov. Also, So the 4<sup>th</sup> moment exists and is finite. Thus,

$$E[X_i^2] \leq (E[|X_i^4|])^{2/4} \leq K^{1/2} < \infty.$$

So the variance of each  $X_i$  is:

$$\text{Var}(X_i) = |E[X_i^2] - (E[X_i])^2| \leq |E[X_i^2]| + (E[X_i])^2,$$

by the triangle inequality (Minkowski). Substituting,

$$\text{Var}(X_i) \leq |E[X_i^2]| + (E[X_i])^2.$$

$$\text{Var}(X_i) \leq K^{1/2} + K^{1/2} = K^*.$$

Now consider the sample mean of the  $X$ 's:

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

Take expectations,

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i].$$

Now define the variance of the sample mean:

$$\begin{aligned}
\text{Var}(\bar{X}) &= E[\bar{X} - E(\bar{X})]^2 \\
&= E\left[\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i]\right]^2 \\
&= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - E[X_i])\right]^2 \\
&= \frac{1}{n^2} E\left[\sum_{i=1}^n (X_i - E[X_i])\right]^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n E[X_i - E[X_i]]^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \\
&\leq \frac{1}{n^2} \sum_{i=1}^n K^* \\
&= \frac{K^*}{n}
\end{aligned}$$

Thus, we have a consistency result from the following:

$$\begin{aligned}
\Pr((\bar{X}_n - E(\bar{X}_n)) \geq \epsilon) &= \Pr((\bar{X}_n - E(\bar{X}_n))^2 \geq \epsilon^2). \\
&\leq \frac{E[(\bar{X}_n - E(\bar{X}_n))^2]}{\epsilon^2}.
\end{aligned}$$

From Chebyshev. Thus,

$$\Pr((\bar{X}_n - E(\bar{X}_n)) \geq \epsilon) \leq \frac{K^*}{n\epsilon^2}.$$

So no matter what the original distribution was of the original  $X$ 's, the means of all those  $X$ 's has a normal distribution and we can use the above inequalities to put bounds on the variance and show that the probability that the sample means will be within  $\epsilon$  of the population mean is bounded. This just the central limit theorem.

## 9.2 Conditional Probability and Stochastic Independence

- **Definition** Consider a probability space  $(\Omega, \mathfrak{A}, P)$ . Let  $B \in \mathfrak{A}$  with  $P(B) > 0$ . Then:

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

The conditional probability of  $A$  given  $B$ .

- Example. 2 coins.  $\Omega = \{HH, HT, TH, TT\}$ . Let  $B = \{HH, HT\}$  and  $A = \{HH\}$ . Thus  $P(A) = 0.25$ . Then:

$$P(A|B) = \frac{1}{2}.$$

$$P(A|B^c) = 0.$$

- Frequency Approach. This definition of conditional probability is easily seen using the frequency approach to probability. Let  $N_B$  be the frequency that  $B$  occurs and  $N_{A \cap B}$  be the frequency that  $A$  and  $B$  occur. Define the relative frequencies as:

$$f_B = \frac{N_B}{N}.$$

$$f_{A \cap B} = \frac{N_{A \cap B}}{N}.$$

Then,

$$f_{A|B} = \frac{N_{A \cap B}}{N_B} = \frac{N_{A \cap B}/N}{N_B/N} = \frac{f_{A \cap B}}{f_B}.$$

- **Theorem 1.** Let  $(\Omega, \mathfrak{A}, P)$  be a probability space. Suppose  $B \in \mathfrak{A}$  with  $P(B) > 0$ . Then the conditional probability function,  $P(\cdot|B) : \mathfrak{A} \mapsto [0, 1]$  defined by  $P(A|B) = P(A \cap B)/P(B)$  for  $A \in \mathfrak{A}$  is a well defined probability measure on  $\mathfrak{A}$ . I.e. it satisfies:

- (1)  $P(\emptyset|B) = 0$ ,  $P(\Omega|B) = 1$ .
- (2)  $P(A|B) \geq 0$  for  $A \in \mathfrak{A}$ .
- (3) If  $A_1, A_2, \dots$  is a sequence of mutually exclusive events in  $\mathfrak{A}$ , then:

$$P\left(\bigcup_{i=1}^{\infty} A_i|B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$

# 10 Lecture 10 - October 5, 2004

## 10.1 More on Conditional Probability

- Recall:

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

- **Theorem 2**

– (1)  $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$ .

- (2) Assume  $A_1, \dots, A_n \in \mathfrak{A}$  and  $P(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$ . Then,

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap \dots \cap A_{n-1}).$$

Or for 3 events:

$$P(A_1 \cap A_2 \cap A_3) = P(A_3|A_1 \cap A_2) * P(A_1 \cap A_2) = P(A_3|A_1 \cap A_2) * P(A_2|A_1) * P(A_1).$$

- **Theorem: The Theorem of Total Probabilities** Suppose  $(\Omega, \mathfrak{A}, P)$  is a probability space. Suppose  $B_1, \dots, B_n$  is a partition of  $\Omega$  (disjoint). Then for every  $A \in \mathfrak{A}$ ,

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

- **Theorem: Bayes Theorem** Suppose  $(\Omega, \mathfrak{A}, P)$  is a probability space. Suppose  $B_1, \dots, B_n$  is a partition of  $\Omega$  (disjoint). Then for every  $A \in \mathfrak{A}$  with  $P(A) > 0$ ,

$$P(B_k|A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A|B_k) * P(B_k)}{\sum_{i=1}^n P(A|B_i) * P(B_i)}.$$

See G-10.1.

- Example. Consider 5 urns with 10 balls in each. Suppose urn  $i$  has  $i$  defective balls. Suppose  $A$  is the event of selecting a defective ball and event  $B_i$  is the selection of a ball from urn  $i$ . What is the probability that you have selected a ball from urn 5 given that it is defective?

$$\begin{aligned} P(B_5|A) &= \frac{P(A|B_5) * P(B_5)}{\sum_{i=1}^5 P(A|B_i) * P(B_i)} = \frac{1/2 * 1/5}{\sum_{i=1}^5 i/10 * 1/5} = \frac{5}{\sum_{i=1}^5 i} = \\ &= \frac{5}{5 * (5 + 1)/2} = \frac{5}{15} = \frac{1}{3}. \end{aligned}$$

- Independence of Events.  $A, B \in \mathfrak{A}$  are independent if:

$$P(A \cap B) = P(A) * P(B).$$

If this is true, then:

$$P(A|B) = P(A), \quad P(B|A) = P(B).$$

But we can only define these conditional probabilities if the event we are conditioning on has positive probability.

- **Definition 3** Suppose there are  $n$  events  $A_i$ . Then the  $A_i$ 's are mutually independent of mutually stochastically independent if:

$$P\left(\bigcap_{i=1}^n A(i)\right) = \prod_{i=1}^n P(A_i),$$

AND all subsets of events have the same property (ie, all sets of 3 events, all pairwise events, etc).

- NOTE! Disjointness does NOT imply independence and vice versa. In fact, if you know that two events are disjoint, you know a LOT about one happening given that the other has happened. For instance, if  $A$  and  $B$  are disjoint and you know that  $A$  has happened, then you also know that  $B$  has NOT happened. Complete dependence.

## 10.2 Marginal and Joint Distributions

- **Theorem 1** Suppose  $X_1$  and  $X_2$  are random variables with  $F(x_1, x_2) = Pr(X_1 \leq x_1, X_2 \leq x_2)$ . Then,

$$F_1(x_1) = Pr(X_1 \leq x_1) = Pr(X_1 \leq x_1, -\infty < x_2 < \infty) = F(x_1, \infty).$$

$F(x_1, x_2)$  is called the joint distribution function and  $F_1(x_1)$  is the marginal distribution function of  $x_1$ .

- For PDFs with a bivariate discrete distribution:

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2).$$

We "SUM-OUT"  $x_2$ . And for a bivariate continuous distribution:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2.$$

We "INTEGRATE-OUT"  $x_2$ .

- Now suppose you have a joint CDF of 5 random variables:

$$F(X_1, X_2, X_3, X_4, X_5).$$

Then,

$$F_*(X_1, X_3) = F(X_1, \infty, X_3, \infty, \infty).$$

- **Definition 2** The moment generating function of  $X = (X_1, \dots, X_n)$  is:

$$M(t_1, t_2, \dots, t_n) = E[e^{t_1 X_1 + t_2 X_2 + \dots + t_n X_n}].$$

So if  $X = (X_1, X_2)$ , then:

$$\left. \frac{\partial^3 M(t_1, t_2)}{\partial t_1^2 \partial t_2} \right|_{t_1=t_2=0} = E[X_1^2 X_2].$$

And,

$$M_1(t_1) = E[e^{t_1 X_1}] = M(t_1, 0).$$

- Note, and we will see this later, if  $X_1$  and  $X_2$  are independent random variables:

$$M(t_1, t_2) = E[e^{t_1 X_1 + t_2 X_2}] = E[e^{t_1 X_1} e^{t_2 X_2}] = E[e^{t_1 X_1}] * E[e^{t_2 X_2}].$$

But ONLY if  $X_1$  and  $X_2$  are truly independent RV's.

- Consider the following example. Suppose  $X = (X_1, X_2)$  and the joint PDF is:

$$f(x_1, x_2) = \begin{cases} 2, & 0 < x_1 < x_2 < 1 \\ 0, & \text{else} \end{cases}$$

See G-10.2. Then the marginal of  $x_1$  is:

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \\ &= \int_{x_1}^1 2 dx_2 \\ &= 2x_2 \Big|_{x_1}^1 \\ &= 2 - 2x_1, \text{ for } 0 < x_1 < 1, \text{ zero else.} \end{aligned}$$

Similarly, the the marginal of  $x_2$  is:

$$\begin{aligned} f_2(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 \\ &= \int_0^{x_2} 2 dx_1 \\ &= 2x_1 \Big|_0^{x_2} \\ &= 2x_2, \text{ for } 0 < x_2 < 1, \text{ zero else.} \end{aligned}$$

See G-10.3 for the marginal PDF of  $x_1$ .

# 11 Lecture 11 - October 7, 2004

## 11.1 Conditional Distributions and Expectations

- Suppose we have 2 discrete random variables,  $X$  and  $Y$ . Then:

$$Pr(X = x|Y = y) = P(\{X = x\}|\{Y = y\}) = \frac{P(\{X = x\} \cap \{Y = y\})}{P(\{Y = y\})} = \frac{f_{x,y}(x, y)}{f_y(y)}.$$

This is denoted:

$$f_{x|y}(x|y),$$

and is called the conditional probability density function. To verify its properties, note  $f_{x|y}(x|y) \geq 0$  and,

$$\sum_x f(x|y) = \sum_x \frac{f(x, y)}{f_y(y)} = \frac{1}{f_y(y)} \sum_x f(x, y) = \frac{1}{f_y(y)} f_y(y) = 1.$$

Good.

- **Definition:** Consider  $X$  and  $Y$  with joint PDF,  $f(x, y)$  and  $f_y(y) > 0$ . Then:

$$f_{x|y}(x|y) = \frac{f_{x,y}(x, y)}{f_y(y)},$$

for both discrete and continuous random variables.

- **Definition:** The conditional CDF is therefore:

$$\text{Discrete: } F_{x|y}(x|y) = \sum_{\{i:x^i \leq x\}} f_{x|y}(x^i|y).$$

$$\text{Continuous: } F_{x|y}(x|y) = \int_{-\infty}^{\infty} f_{x|y}(u|y) du.$$

We use  $u$  as the integration variable here for clarity.

- **Definition:** Conditional probability of an event:

$$P(X \in B|Y = y) = \int_B f(x|y) dx.$$

If  $B = [a, b]$ , then:

$$Pr(a \leq x \leq b|Y = y) = \int_a^b f_{x|y}(x|y) dx.$$

- **Definition:** Given two RVs,  $X$  and  $Y$  and a function  $g(x, y)$ , then:

$$E[g(x, y)|Y = y] = \int_{-\infty}^{\infty} g(x, y) f_{x|y}(x|y) dx.$$

If  $g(x, y) = u(x)$ ,

$$E[u(x)|Y = y] = \int_{-\infty}^{\infty} u(x)f_{x|y}(x|y)dx.$$

And if  $u(x) = x$ ,

$$E[x|Y = y] = \int_{-\infty}^{\infty} xf_{x|y}(x|y)dx.$$

- **Definition:** Conditional Variance:

$$\begin{aligned} E\left[\underbrace{X - E(X|Y = y)}_{\mu_{x|y}}\right]^2|Y = y] &= \int_{-\infty}^{\infty} [x - \mu_{x|y}]^2 f_{x|y}(x|y)dx. \\ &= E[X^2|Y = y] - \left(E[X|Y = y]\right)^2 \end{aligned}$$

- Recall our example from last lecture: Suppose  $X = (X_1, X_2)$  and the joint PDF is:

$$f(x_1, x_2) = \begin{cases} 2, & 0 < x_1 < x_2 < 1 \\ 0, & \text{else} \end{cases}$$

We found the marginals:

$$f_1(x_1) = \begin{cases} 2(1 - x_1), & 0 < x_1 < 1 \\ 0, & \text{else} \end{cases}$$

$$f_2(x_2) = \begin{cases} 2x_2, & 0 < x_2 < 1 \\ 0, & \text{else} \end{cases}$$

Thus, we can find the conditional PDF as well:

$$f(x_1|x_2) = \begin{cases} \frac{f(x_1, x_2)}{f_2(x_2)} = \begin{cases} \frac{2}{2x_2} & 0 < x_1 < x_2 \\ 0 & \text{else} \end{cases} & 0 < x_2 < 1 \\ \text{Undefined} & x_2 \notin (0, 1) \end{cases}$$

Similarly,

$$f(x_2|x_1) = \begin{cases} \frac{f(x_1, x_2)}{f_1(x_1)} = \begin{cases} \frac{2}{2(1 - x_1)} & x_1 < x_2 < 1 \\ 0 & \text{else} \end{cases} & 0 < x_1 < 1 \\ \text{Undefined} & x_1 \notin (0, 1) \end{cases}$$

And the conditional expected value:

$$E[x_1|x_2] = \begin{cases} \int_{-\infty}^{\infty} x_1 f(x_1|x_2) dx_1 = \int_0^{x_2} x_1 \frac{2}{2x_2} = \frac{x_2}{2} & 0 < x_2 < 1 \\ \text{Undefined} & x_2 \notin (0, 1) \end{cases}$$

Notice the expectation depends on  $x_2$ . See graph G-10.2 for an illustration. As  $x_2$  approaches 0, the range of expectation for  $x_1$  gets smaller (as does its conditional variance). Thus conditional expectations will usually depend on the conditioning variable:

$$E[u(x)|y] = h(y).$$

Thus  $h(Y) = E[u(x)|Y]$  is a new function of the conditioning variable  $Y$ .

- **Theorem:** Law of Iterated Expectations.

$$E_Y[E[g(X)|Y]] = E[g(X)].$$

And for  $g(X) = X$ ,

$$E[E[X|Y]] = E[X].$$

Proof: Consider the conditional PDF,  $f(x|y) = f(x, y)/f_y(y)$ . Then:

$$h(y) = E[g(x)|y] = \int_{-\infty}^{\infty} g(x) f(x|y) dx.$$

Take expectations:

$$\begin{aligned} E[h(y)] &= E[E[g(x)|y]] = \int h(y) f_y(y) dy \\ &= \int \left[ \int g(x) \frac{f(x, y)}{f_y(y)} dx \right] f_y(y) dy \\ &= \int \frac{1}{f_y(y)} \left[ \int g(x) f(x, y) dx \right] f_y(y) dy \\ &= \int \int g(x) f(x, y) dx dy \\ &= \int \int g(x) f(x, y) dy dx \\ &= \int g(x) \left[ \int f(x, y) dy \right] dx \\ &= \int g(x) f_x(x) dx \\ &= E[g(x)] \end{aligned}$$

- One interesting application is regression analysis where say you have a model:  $y_t =$

$x_t\beta + u_t$  and you know  $E[u_t|x_t] = 0$ . Then by the previous theorem:

$$E[u_t] = E[E(u_t|x_t)] = E[0] = 0.$$

Good.

## 11.2 Stochastic Independence and Random Variables

- Suppose we have two random variables,  $X$  and  $Y$  with conditional density,  $f(x|y)$ . Then  $X$  is independent of  $Y$  if:

$$f(x|y) = f_x(x).$$

This implies:

$$f(x|y) = \frac{f(x, y)}{f_y(y)} \implies f(x, y) = f_x(x) * f_y(y).$$

- Now consider the probability of  $X_1$  and  $X_2$  being outcomes of two events,  $B_1$  and  $B_2$ :

$$\begin{aligned} P(X_1 \in B_1, X_2 \in B_2) &= \int_{B_1} \int_{B_2} f(x_1, x_2) dx_2 dx_1 \\ &= \int_{B_1} \int_{B_2} f_1(x_1) f_2(x_2) dx_2 dx_1 \\ &= \int_{B_1} f_1(x_1) dx_1 * \int_{B_2} f_2(x_2) dx_2 \\ &= P(X_1 \in B_1) * P(X_2 \in B_2). \end{aligned}$$

- If  $B_1 = (-\infty, x_1]$  and  $B_2 = (-\infty, x_2]$ , then,

$$P(X_1 \in B_1, X_2 \in B_2) = F(x_1, x_2).$$

And,

$$P(X_1 \in B_1) = F_1(x_1).$$

$$P(X_2 \in B_2) = F_2(x_2).$$

It can be shown that this also holds for all Borel sets.

## 12 Lecture 12 - October 12, 2004

### 12.1 More on Stochastic Independence

- **Theorem:** A necessary and sufficient condition for independence of random variables is the existence of non-negative functions,  $g_1(x_1)$  and  $g_2(x_2)$  such that:

$$f(x_1, x_2) = g_1(x_1)g_2(x_2) \quad \forall (x_1, x_2) \in \mathfrak{R}^2.$$

Note that  $g_i(x_i)$  could be the marginal PDF of  $x_i$ . The important part of this theorem is that it must hold for all  $x_1, x_2$  pairs in the real plane. If it doesn't then we don't have independence.

- Example. Consider 2 random variables,  $(X_1, X_2)$  with pdf:

$$f(x_1, x_2) = \begin{cases} 8x_1x_2, & 0 < x_1 < x_2 < 1 \\ 0, & \text{else} \end{cases}$$

Define  $g_1(x_1) = 8x_1$  and  $g_2(x_2) = x_2$ . See graph G-12.1 for the support of this distribution. Take a point  $x_1 = 3/4$  and  $x_2 = 1/2$ . Clearly (from the graph) this has density 0 but:

$$f(x_1, x_2) = g_1(x_1)g_2(x_2) = 8x_1x_2 = 8(3/4)(1/2) = 3 \neq 0.$$

- Thus the "For all" portion of the theorem is critical. As long as the support of the variables in  $\mathfrak{R}^2$  is not a rectangle, there will be some sort of dependence. If it is a rectangle, then we can't say for sure if  $X_1$  and  $X_2$  are independent - we would have to check.
- **Theorem:** Suppose  $X_1$  and  $X_2$  are independent RVs. Define:

$$Y_1 = u_1(X_1), \quad Y_2 = u_2(X_2).$$

Where  $u_1$  and  $u_2$  are measurable functions. Then  $Y_1$  and  $Y_2$  are also independent. "Functions of independent random variables are independent."

- Example. Setup as in theorem. Then the event  $\{Y_1 \in B_1\} = \{u_1(X_1) \in B_1\} = \{X_1 \in \underbrace{u_1^{-1}(B_1)}_{A_1}\}$ . Since  $u_1$  is measurable and  $B_1$  is a borel set, then  $u_1^{-1}(B_1)$  is also a borel set. Thus,

$$\begin{aligned} Pr(Y_1 \in B_1, Y_2 \in B_2) &= P(X_1 \in A_1, X_2 \in A_2) \\ &= P(X_1 \in A_1) * P(X_2 \in A_2) \\ &= P(Y_1 \in B_1) * P(Y_2 \in B_2) \end{aligned}$$

So  $Y_1$  and  $Y_2$  are independent.

- **Theorem:** Suppose  $X_1$  and  $X_2$  are independent RVs. If  $E[X_1] < \infty$  and  $E[X_2] < \infty$ , then  $E[X_1X_2] < \infty$  and:

$$E[X_1X_2] = E[X_1]E[X_2].$$

Proof:

$$\begin{aligned} E[X_1X_2] &= \int \int x_1x_2f(x_1, x_2)dx_1dx_2 = \int \int x_1x_2f_1(x_1)f_2(x_2)dx_1dx_2 = \\ &= \int x_1f_1(x_1)dx_1 * \int x_2f_2(x_2)dx_2 = E[X_1] * E[X_2]. \end{aligned}$$

- **Theorem:** The Moment Generating Function under independence. Suppose  $X_1$  and  $X_2$  are independent. Then:

$$M(t_1, t_2) = E[e^{t_1X_1+t_2X_2}] = E[e^{t_1X_1}e^{t_2X_2}] = E[e^{t_1X_1}] * E[e^{t_2X_2}] = M_1(t_1) * M_2(t_2).$$

This also holds in reverse: If the moment generating function is separable like we have here, then  $X_1$  and  $X_2$  are independent.

- **Theorem:** If  $X$  and  $Y$  are independent random variables:

$$E[X|Y = y] = E[X].$$

## 12.2 Covariance and the Correlation Coefficient

- **Definition:** Covariance:

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x\mu_y.$$

A positive covariance means that points generally are positively related and vice versa. Covariance depends on units of the variable so hence we will soon define the correlation coefficient.

- If  $Y = X$ ,  $\sigma_{xx} = \sigma_x^2 = \text{variance}$ .
- A sufficient condition for the existence of the covariance is for the variance of  $x$  and the variance of  $y$  to exist. You can see this using the Holder Inequality:

$$\begin{aligned} [E[(X - \mu_x)(Y - \mu_y)]]^2 &= |E[(X - \mu_x)(Y - \mu_y)]|^2 \\ &\leq [E|(X - \mu_x)(Y - \mu_y)|]^2 \\ &\leq E(X - \mu_x)^2 E(Y - \mu_y)^2 \\ &= \sigma_x^2 \sigma_y^2 < \infty \end{aligned}$$

- **Definition:** The correlation coefficient:

$$\text{Corr}(X, Y) = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Given  $0 < \sigma_x, \sigma_y < \infty$ .

- Example. Consider a random variable  $X$  with pdf:

$$f(x) = \begin{cases} \frac{1}{2}, & x = +1, -1 \\ 0, & \text{else} \end{cases}$$

So  $E[X] = 0$ . Now let  $Y = X^2$ . So:

$$\begin{aligned} \sigma_{xy} &= E[XY] - \mu_x \mu_y. \\ &= E[X * X^2] - 0\mu_y. \\ &= E[X^3]. \\ &= (+1)^3 * (1/2) + (-1)^3 * (1/2) = 0. \end{aligned}$$

So there is clear dependence between  $X$  and  $Y$  though the covariance is 0. (The relationship is Non-Linear so what the covariance is telling us is that there is no Linear relationship).

## 13 Lecture 13 - October 14, 2004

### 13.1 More on Covariance and Correlation

- Consider two RV's,  $X$  and  $Y$ . Recall the covariance is:

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)].$$

Define  $Z = a + bX$ . Then  $E[Z] = \mu_z = a + b\mu_x$ . So  $Z - \mu_z = a + bX - a - b\mu_x = b(X - \mu_x)$ . Thus,

$$\sigma_{zy} = E[(Z - \mu_z)(Y - \mu_y)] = E[(b(X - \mu_x)(Y - \mu_y))] = b\sigma_{xy}.$$

Thus additive constants do not change the covariance but scaling does. Also,

$$\sigma_z^2 = E[(Z - \mu_z)^2] = b^2 E[(X - \mu_x)^2] = b^2 \sigma_x^2.$$

- Define the correlation coefficient as:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

So,

$$\rho_{zy} = \frac{\sigma_{zy}}{\sigma_z \sigma_y} = \frac{b\sigma_{xy}}{|b|\sigma_x \sigma_y} = \frac{b}{|b|} \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \text{sign}(b) * \rho_{xy}.$$

So scaling a RV only effects  $\rho$  in the sign. If  $b < 0$ , the sign will switch and vice versa.  $\rho$  is independent of units of measurement.

- Note that  $|\rho_{xy}| \leq 1$ .
- If  $Y = a + bX$ ,  $\rho_{xy} = \text{sign}(b) * 1$ .

### 13.2 Special Distributions

#### Gamma Distribution

- Consider the gamma function:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy, \quad \alpha > 0.$$

The gamma function satisfies:

- (1)  $\Gamma(1) = 1$ .
- (2)  $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$  for  $\alpha > 1$ .
- (3)  $\Gamma(\alpha) = (\alpha - 1)(\alpha - 2) \dots (2)(1)\Gamma(1) = (\alpha - 1)!$  for  $\alpha = 1, 2, 3, \dots$

- Now let  $y = x/\beta$  with  $\beta > 0$ . Substituting,

$$\Gamma(\alpha) = \int_0^\infty \frac{x^{\alpha-1}}{\beta^{\alpha-1}} e^{-x/\beta} \underbrace{\left| \frac{dy}{dx} \right|}_{1/\beta} dx, \quad \alpha, \beta > 0.$$

Note we need to multiply by the jacobian to do this change of variables. Thus,

$$\Gamma(\alpha) = \int_0^\infty \frac{x^{\alpha-1}}{\beta^\alpha} e^{-x/\beta} dx, \quad \alpha, \beta > 0.$$

And finally divide both sides by  $\Gamma(\alpha)$ , yields:

$$1 = \int_0^\infty \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta} dx, \quad \alpha, \beta > 0.$$

So since  $\Gamma(\alpha) > 0$  and the whole thing integrates to 1, we have the PDF for  $X$  defined as:

$$f_x(X) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta}, \quad \alpha, \beta > 0.$$

See G-13.1 for a few pictures of the gamma distribution. It is useful for waiting times, decay rates, time until death, etc.

- Properties:

- $E[X] = \alpha\beta$ .
- $\sigma_x^2 = \alpha\beta^2$ .
- MGF:  $M(t) = \frac{1}{(1 - \beta t)^\alpha}$ ,  $t < 1/\beta$ .

### Exponential Distribution (Special Case of Gamma)

- Start with gamma PDF:

$$f_x(X) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta}, \quad \alpha, \beta > 0.$$

Let  $\alpha = 1$  and  $\beta = 1/\lambda$ . Thus, the exponential PDF:

$$f_x(X) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty.$$

- Properties:

- $E[X] = \frac{1}{\lambda}$ .
- $\sigma_x^2 = \frac{1}{\lambda^2}$ .

– MGF:  $M(t) = \frac{\lambda}{\lambda - t}, t < \lambda.$

See G-13.2 for a picture of the exponential distribution.

### Chi-Squared Distribution (Special Case of Gamma)

- Start with gamma PDF:

$$f_x(X) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta}, \quad \alpha, \beta > 0.$$

Let  $\alpha = \frac{r}{2}$  and  $\beta = 2$ . Thus, the  $\chi^2$  PDF:

$$f_x(X) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{\frac{r}{2}-1} e^{-x/2}, \quad 0 \leq x < \infty.$$

We say that  $X \sim \chi^2(r)$  where  $r$  is the degrees of freedom.

- Properties:

- $E[X] = r.$
- $\sigma_x^2 = 2r.$
- MGF:  $M(t) = (1 - 2t)^{-r/2}, t < 1/2.$

- Note the term: “degrees of freedom” comes from regression analysis where if you estimate the regression equation:

$$y_t = X_t\beta + u_t, \quad u_t \sim iid N(0, 1), \quad t = 1, \dots, n.$$

Then,

$$\sum_{t=1}^n \frac{u_t^2}{\sigma^2} = \sum_{t=1}^n u_t^2 \sim \chi^2(n).$$

But suppose we estimate  $\hat{\beta}$ , a  $k \times 1$  vector. Then,

$$y_t = X_t\hat{\beta} + \hat{u}_t,$$

and:

$$\sum_{t=1}^n \hat{u}_t^2 \sim \chi^2(n - k).$$

Since we used  $k$  observations to estimate  $\hat{\beta}$ , we have essentially LOST  $k$  degrees of freedom. And this is where the expression comes from.

## Normal or Gaussian Distribution

- PDF:

$$f_x(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(X-\mu_x)^2}.$$

We say  $X \sim N(\mu, \sigma^2)$  where  $\mu$  and  $\sigma^2$  are just parameters. Utilizing the MGF below, you find that these parameters are exactly the mean and variance of a normally distributed RV.

- Properties:

- $E[X] = \mu \in \mathfrak{R}$ .
- $\sigma_x^2 = \sigma^2 > 0$ .
- MGF:  $M(t) = e^{\mu t + (1/2)\sigma^2 t^2}$ .

Clearly  $f(x) \geq 0$  for all  $x$  because  $\exp(\cdot)$  and  $\sigma > 0$ , but to show that it integrates to 1 is a bit trickier. Next time...

# 14 Lecture 14 - October 19, 2004

## 14.1 Integration of the Normal

- Consider the gaussian PDF:

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

To show this function integrates to 1, ie:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1,$$

consider the following integral:

$$I = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy.$$

Square it:

$$I^2 = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz.$$

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2}} dydz.$$

Switch to polar coordinates such that:

$$y = r \cos(\theta), \quad z = r \sin(\theta).$$

And the jacobian of the tranformation:

$$J = \begin{bmatrix} \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -r \sin(\theta) \\ \sin(\theta) & r \cos(\theta) \end{bmatrix}.$$

Thus  $|J| = r * (\cos(\theta))^2 + r * (\sin(\theta))^2 = r$ . Also note that  $y^2 + z^2 = r^2 * (\cos)^2 + r^2 * (\sin)^2 = r^2$ . Substitute in for  $y$  and  $z$  and rewrite the integral:

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} r dr d\theta.$$

Consider the inner integral and let  $x = r^2/2$ . So:

$$\frac{dx}{dr} = r \implies dx = r dr.$$

Substitute again:

$$I^2 = \int_0^{2\pi} \underbrace{\int_0^\infty e^{-x} dx}_{1} d\theta.$$

$$I^2 = \int_0^{2\pi} d\theta = 2\pi.$$

$$I = \int_{-\infty}^\infty e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}.$$

And of course:

$$\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 1.$$

So, let  $y = (x - \mu)/\sigma$  so that  $dy/dx = \frac{1}{\sigma}$  and  $dy = \frac{1}{\sigma} dx$ . Substituting in:

$$\int_{-\infty}^\infty \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \frac{1}{\sigma}}_{N(x)} dx = 1.$$

So the PDF of  $x$ ,  $N(x)$  does indeed integrate to 1.

## 14.2 Moment Generating Function of the Normal Distribution

- Consider the MGF of the normal:

$$\begin{aligned}
 M(t) = E[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{tx - \frac{1}{2\sigma^2}(x-\mu)^2} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[-2tx\sigma^2 + (x-\mu)^2]} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[-2tx\sigma^2 + x^2 - 2x\mu + \mu^2]} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[x^2 - 2x(t\sigma^2 + \mu) + \mu^2]} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[x^2 - 2x(t\sigma^2 + \mu) + \mu^2 + 2\sigma^2\mu t - \sigma^4 t^2 - 2\sigma^2\mu t - \sigma^4 t^2]} dx. \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[[x^2 - (t\sigma^2 + \mu)]^2 - 2\sigma^2\mu t - \sigma^4 t^2]} dx. \\
 &= e^{-\frac{1}{2\sigma^2}[-2\sigma^2\mu t - \sigma^4 t^2]} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\underbrace{[x^2 - (t\sigma^2 + \mu)]^2}_{\mu^*}} dx. \\
 &= e^{\mu t + 1/2\sigma^2 t^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\mu^*]} dx}_{1}. \\
 &= e^{\mu t + 1/2\sigma^2 t^2}.
 \end{aligned}$$

So this is the MGF of the normal and  $X \sim N(\mu, \sigma^2)$ .

- Now suppose  $X \sim (\mu, \sigma^2)$ . Let:

$$W = \frac{X - \mu}{\sigma}.$$

$W \sim (0, 1)$ , standardized. Note:

$$E[W] = \frac{E[X] - \mu}{\sigma} = 0.$$

And,

$$\sigma_w^2 = E[W^2] = \frac{1}{\sigma^2} E[(X - \mu)^2] = \frac{\sigma^2}{\sigma^2} = 1.$$

- If  $X \sim N(\mu, \sigma^2)$ , then  $W = (X - \mu)/\sigma \sim N(0, 1)$ . This is only true for the Normal. It is usually NOT the case.

Proof: Consider the MGF of W:

$$\begin{aligned} M_w(t) = E[e^{tW}] &= E\left[e^{t\left(\frac{x - \mu}{\sigma}\right)}\right] \\ &= E\left[e^{-\frac{t\mu}{\sigma} + \frac{tx}{\sigma}}\right] \\ &= E\left[e^{-\frac{t\mu}{\sigma}} e^{\frac{tx}{\sigma}}\right] \\ &= e^{-\frac{t\mu}{\sigma}} E\left[e^{\frac{tx}{\sigma}}\right] \\ &= e^{-\frac{t\mu}{\sigma}} E[e^{sx}], \quad (s = t/\sigma) \\ &= e^{-\frac{t\mu}{\sigma}} e^{\mu s + 1/2 * \sigma^2 s^2} \\ &= e^{-\frac{t\mu}{\sigma}} e^{\frac{\mu t}{\sigma} + 1/2 * \sigma^2 \frac{t^2}{\sigma^2}} \\ &= e^{-\frac{t\mu}{\sigma}} e^{\frac{\mu t}{\sigma} + 1/2 * t^2} \\ &= e^{-\frac{t\mu}{\sigma} + \frac{\mu t}{\sigma}} e^{1/2 * t^2} \\ &= e^0 e^{1/2 * t^2} \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

And this last equation is equivalent to the MGF of the normal setting  $\mu = 0$  and  $\sigma = 1$ . So  $W \sim N(0, 1)$  because there is a one to one correspondence between the MGF and the distributions.

- If you have a normal RV, you can always standardize it to use the lookup tables to find the correct probabilities.
- The CDF of the standard normal:

$$N(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

So  $Pr(c_1 < X < c_2) = N(c_2) - N(c_1)$  if  $X$  is standard normal.

- If  $W \sim N(0, 1)$  then  $W^2 \sim \chi^2(1)$ .

## 14.3 Distributions of Functions of Random Variables

### Sampling Theory

- Suppose  $X_i$  is a RV  $\sim (\mu, \sigma^2)$ ,  $i = 1 \dots n$ .

- Let:

$$Y = \frac{1}{n} \sum_{i=1}^n X_i.$$

$Y$  is a statistic since it does not depend on unknown parameters.

- **Definition:** Random Sample. If  $X_1, \dots, X_n$  are independently and identically distributed RVs all with CDF,  $F(X)$ , then:

$$F_*(X_1, \dots, X_n) = F(X_1)F(X_2) \dots F(X_n) = \prod_{i=1}^n F(X_i).$$

- Suppose we have random variables:  $Z_1, \dots, Z_n$ , with distribution function:

$$F(Z_1, \dots, Z_n, \theta).$$

Where  $\theta$  is an unknown set of parameters (could be regression coefficients, variance of the error term, etc).

- If  $Z_i \sim \text{iid } N(\mu, \sigma^2)$  and  $Y = \frac{1}{n} \sum Z_i$ , call  $Y$  the sample mean of  $X$ . Note  $E[Y] = \mu$  and  $Var(Y) = \frac{1}{n} \sigma^2$ . We would like to find functions of the  $Z$ 's that give us information about  $\theta$ .  $Y$  is a good choice (notice it is a statistic). The sample variance is also a good choice.

# 15 Lecture 15 - October 21, 2004

## 15.1 Sample Mean and Variance

- Consider a random sample:  $X_1, \dots, X_n$ . Then:

$$E[X_i] = \mu \forall i.$$

$$Var[X_i] = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \sigma^2 \forall i.$$

- Consider two statistics:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \text{Sample Mean.}$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \equiv \text{Sample Variance.}$$

- For random samples, we assume that the mean and variance are the same for each  $X_i$ . Suppose not. Suppose  $E[X_i] = \mu_i$ , different for each  $i$ . Then the “Expected” sample mean is:

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu_i,$$

which is the average of the population means. If all population means are the same, a true random sample, then  $E[\bar{X}] = \mu$ . So the sample mean would be an unbiased estimator of the population mean.

- Now consider the variance of the sample mean when the random variables have different

means and variances:

$$\begin{aligned}
 \text{Var}(\bar{X}) = E(\bar{X} - E(\bar{X}))^2 &= E\left[\left(\frac{1}{n} \sum X_i - \frac{1}{n} \sum \mu_i\right)^2\right] \\
 &= E\left[\left(\frac{1}{n} \sum X_i - \mu_i\right)^2\right] \\
 &= \frac{1}{n^2} \left[ \sum_i \sum_j \underbrace{E(X_i - \mu_i)(X_j - \mu_j)}_{\sigma_{ij}} \right] \\
 &= \frac{1}{n^2} \sum_i \sum_j \sigma_{ij} \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 + \frac{1}{n^2} \underbrace{\sum_i \sum_{j \neq i} \sigma_{ij}}_{\text{covariance terms}} \\
 &= \frac{n * \sigma^2}{n^2} = \frac{\sigma^2}{n} \text{ if Random Sample!!}
 \end{aligned}$$

So those covariance terms are all zero if we have an iid random sample, but otherwise, they must be added in.

- So we have a random sample with mean  $\mu$  and variance  $\sigma^2$  and the sample mean of that random sample has expectation  $\mu$  and variance  $\sigma^2/n$ .
- Now suppose  $\mu$  is known but  $\sigma^2$  is unknown. Let  $Z_i = (X_i - \mu)^2$ . Thus,

$$E[Z_i] = E(X_i - \mu)^2 = \sigma^2.$$

So,

$$\frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \hat{\sigma}^2.$$

So if  $\mu$  is known, we can estimate the variance using this statistic.

- Now suppose  $\mu$  is unknown. We could try to use the following sample variance:

$$S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2.$$

Is it UNBIASED? NO. Consider the following:

$$E[(X_i - \bar{X})^2] = E(X_i^2) - 2E(X_i \bar{X}) + E(\bar{X}^2).$$

First term: Recall that  $\sigma^2 = E[X_i^2] - \mu^2$  so:

$$E[X_i^2] = \sigma^2 + \mu^2.$$

Third term: By the same reasoning (shown above):

$$E[\bar{X}^2] = \sigma^2/n + \mu^2.$$

Second term:

$$\begin{aligned} E[X_i \bar{X}] &= E[X_i * \frac{1}{n} \sum_{i=1}^n X_i] \\ &= \frac{1}{n} E[X_i(X_1 + X_2 + \dots + X_n)] \\ &= \frac{1}{n} E[X_i X_1 + X_i X_2 + \dots + X_i^2 + \dots + X_i X_n] \\ &= \frac{1}{n} \underbrace{E[X_i X_1]}_{\mu^2} + \underbrace{E[X_i X_2]}_{\mu^2} + \dots + \underbrace{E[X_i^2]}_{\sigma^2 + \mu^2} + \dots + \underbrace{E[X_i X_n]}_{\mu^2} \\ &= \frac{1}{n} [(n-1) * \mu^2 + (\sigma^2 + \mu^2)] \\ &= \frac{1}{n} [n\mu^2 + \sigma^2] \\ &= \mu^2 + \frac{\sigma^2}{n} \end{aligned}$$

Note that  $E[X_i X_j] = E[X_i]E[X_j]$  because the  $X$ 's are independent.  
Thus plugging in:

$$\begin{aligned} E[S^2] &= \frac{1}{n} \sum E[(X_i - \bar{X})^2] \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{[\sigma^2 + \mu^2]}_{\text{First}} \underbrace{- 2(\mu^2 + \frac{\sigma^2}{n})}_{\text{Second}} \underbrace{+ \frac{\sigma^2}{n} + \mu^2}_{\text{Third}} \\ &= \frac{1}{n} \sum_{i=1}^n [\sigma^2 + \mu^2 - 2\mu^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n} + \mu^2] \\ &= \frac{1}{n} \sum_{i=1}^n [\sigma^2 - \frac{\sigma^2}{n}] \\ &= \frac{1}{n} [n\sigma^2 - \sigma^2] \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n\sigma^2 - \sigma^2}{n} = \frac{n-1}{n} \sigma^2 \neq \sigma^2 \end{aligned}$$

So  $S^2$  is a biased estimator of  $\sigma^2$  !!

- Consider a slightly different statistic:

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Note that

$$\frac{n}{n-1} S^2 = \frac{n}{n-1} \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 = S_1^2$$

So:

$$E[S_1^2] = \frac{n}{n-1} * E[S^2] = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2.$$

So  $S_1^2$  is unbiased! The reason is that we lose a degree of freedom when estimating  $\bar{X}$  so we divide by  $n-1$ .

## 15.2 Technique 1: Cumulative Distribution Function Technique

- In this section we are concerned with determining the distribution of functions of random variables. The first is called the CD function technique.
- Consider the following example.  $X$  is a continuous RV with density:

$$f(x) = \begin{cases} \frac{1}{2}, & -1 \leq x \leq 1 \\ 0, & \text{else} \end{cases}$$

Now consider a RV  $Y$  such that  $Y = X^2$ . What is the distribution of  $Y$ ? Consider two cases:

- If  $y < 0$ :

$$G(y) = Pr(Y \leq y) = Pr(X^2 \leq y) = 0.$$

- If  $y \geq 0$ , G-15.1:

$$\begin{aligned} G(y) &= Pr(Y \leq y) = Pr(X^2 \leq y) = Pr(|X| \leq \sqrt{y}) = Pr(-\sqrt{y} \leq x \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx. \\ &= \begin{cases} \int_{-1}^1 1/2 dy = 1, & y > 1 \\ \int_{-\sqrt{y}}^{\sqrt{y}} 1/2 dy = \sqrt{y}, & 0 \leq y \leq 1 \end{cases} \end{aligned}$$

- So the CDF of  $Y$  is:

$$G(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}, & 0 \leq y \leq 1 \\ 1, & y > 1 \end{cases}$$

- Differentiating with respect to  $y$  gives us the PDF of  $y$ :

$$g(y) = \begin{cases} 1, & y < 0 \\ \frac{1}{2\sqrt{y}}, & 0 \leq y \leq 1 \\ 0, & y > 1 \end{cases}$$

- **Theorem:** The Probability Integral Transform. Let  $y = F(x)$ , the CDF of  $x$ , so that the range of  $y$  is  $[0, 1]$ . Then  $y \sim U[0, 1]$ . If we were to start with a uniform  $Y$  and then apply  $X = F^{-1}(Y)$ , this will yield  $X$  with distribution  $F(X)$ . This method is very useful for generating distributions of random variables. As long as we have a good random number generator to generate uniform RVs, we can generate a random sample from any distribution buy simply applying this theorem.

### 15.3 Technique 2: The Transformation Technique (Discrete)

- Suppose a discrete RV,  $X \sim \text{poisson}$  with PDF:

$$f(x) = \begin{cases} \frac{e^{-\mu} \mu^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{else} \end{cases}$$

- Let  $\mathfrak{X} = \{x : f(x) > 0\}$ , ie the support of the distribution,  $\mathfrak{X} = \{x : x = 0, 1, 2, \dots\}$ . Now define another RV:

$$Y = u(X) = 4X, \quad y = 4x.$$

And define the following set:

$$\mathcal{Y} = \{y : y = u(x), x \in \mathfrak{X}\} = \{y : y = 0, 4, 8, 12, \dots\}.$$

Note that  $x = w(y) = \frac{1}{4}y$ . Then the PDF of  $Y$  is as follows:

$$g(y) = Pr(Y = y) = \begin{cases} Pr(x = \frac{1}{4}y) = \frac{e^{-\mu} \mu^{y/4}}{(y/4)!}, & y = 0, 4, 8, 12, \dots \\ 0, & \text{else} \end{cases}$$

Or more simply:

$$g(y) = \begin{cases} f(w(y)), & y \in \mathcal{Y} \\ 0, & \text{else} \end{cases}$$

where  $w(y) = \frac{1}{4}y$ .

- Note that in this example we have a 1:1 mapping between  $X$  and  $Y$ . It could be the case that the mapping is not 1:1. If two different  $X$ 's map to the same  $Y$ , then  $g(y) = f(x^1) + f(x^2)$ .

## 16 Lecture 16 - October 26, 2004

### 16.1 Change of Variables Technique for Continuous RVs

- Suppose  $X$  is a continuous RV with PDF:

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

Thus

$$\mathfrak{X} = \{x : f(x) > 0\} = \{x : 0 < x < 1\}.$$

Now, let  $y = u(x) = 8x^3$ . Thus,

$$x = w(y) = \frac{1}{2}y^{1/3}.$$

Since  $u : \mathfrak{X} \mapsto \mathcal{Y}$ ,

$$\mathcal{Y} = \{y : y = u(x), x \in \mathfrak{X}\} = \{y : 0 < y < 8\}.$$

- So consider the following probability:

$$\begin{aligned} Pr(a < Y < b) &= Pr(a < 8x^3 < b) \\ &= Pr\left(\frac{1}{2}a^{1/3} < x < \frac{1}{2}b^{1/3}\right) \\ &= \int_{0.5a^{1/3}}^{0.5b^{1/3}} 2x dx \\ &= \int_a^b 2\left(\frac{1}{2}y^{1/3}\right)|J|dy \end{aligned}$$

$$\text{Where : } x = \frac{1}{2}y^{1/3}$$

$$\begin{aligned} \text{And : } |J| &= x'(y) = \frac{1}{6}y^{-2/3} \\ &= \int_a^b 2\left(\frac{1}{2}y^{1/3}\right)\frac{1}{6y^{2/3}}dy \\ &= \int_a^b \frac{1}{6y^{1/3}}dy \end{aligned}$$

So since we defined this for an arbitrary  $a$  and  $b$ ,  $\frac{1}{6y^{1/3}}$  must be the PDF of  $Y$ .

- In general however,

$$g(y) = f(w(y)) * \left| \frac{\partial w(y)}{\partial y} \right| \text{ for } y \in \mathcal{Y}.$$

So to use this technique on a continuous random variable, the only thing different from the discrete case is that we multiply by the absolute value of the determinant of the jacobian.

- **Theorem:** Suppose  $X_1, \dots, X_n$  are continuous RVs and define:

$$Y_1, \dots, Y_n \ni Y_i = u_i(X_1, \dots, X_n).$$

Let  $\mathfrak{X} = \{(X_1, \dots, X_n) : f(X_1, \dots, X_n) > 0\}$ . If  $y_i = u_i(x_1, \dots, x_n)$  then  $x_i = w_i(y_1, \dots, y_n)$ . And:

$$\mathcal{Y} = \{(y_1, \dots, y_n) : y_i = u_i(x_1, \dots, x_n), (x_1, \dots, x_n) \in \mathfrak{X}\}.$$

Define the jacobian matrix as:

$$J = \begin{bmatrix} \frac{\partial w_1}{\partial y_1} & \cdots & \frac{\partial w_n}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_n}{\partial y_n} & \cdots & \frac{\partial w_n}{\partial y_n} \end{bmatrix}.$$

Thus,

$$g(y_1, \dots, y_n) = \begin{cases} f(w_1(y_1, \dots, y_n), \dots, w_n(y_1, \dots, y_n)) |J|, & (y_1, \dots, y_n) \in \mathcal{Y} \\ 0, & \text{else} \end{cases}$$

- If the mapping between the old variables and the new is not 1:1, then split the sample into two disjoint sets which are 1:1 and add the probabilities. See notes.

## 16.2 Student $t$ and $F$ Distribution

### Student $t$

- A random variable  $X$  with density :

$$f(x) = \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r} \Gamma(r/2)} (1 + (x^2/r))^{-(r+1)/2}, \quad -\infty < x < \infty,$$

is distributed  $t(r)$ . As  $r \rightarrow \infty$ ,  $t(r) \rightarrow N(0, 1)$ .

- The  $t$  has finite moments ( $E|X|^s$ ) of order  $s$  for all  $s < r$ .  $E[X] = \mu_x = 0$  if  $r > 1$ , otherwise the mean does not exist.  $Var[X] = \sigma_x^2 = \frac{r}{r-2}$  if  $r > 2$ .
- If  $r = 1$ , the distribution is called the Cauchy Distribution.

- **Theorem:** If  $Z \sim N(0, 1)$  and  $Y_2 \sim \chi^2(r)$ , with  $Z$  and  $Y_2$  independent. Then:

$$X = \frac{Z}{\sqrt{Y_2/r}} \sim t(r).$$

- Consider an OLS regression with normally distributed error.  $Z = (\hat{\beta} - \beta)/\sigma_{\hat{\beta}} \sim N(0, 1)$  and  $Y_2/r = \frac{\hat{\sigma}_{\hat{\beta}^2}}{\sigma_{\hat{\beta}^2}} \sim \chi^2(r)$ , then:

$$X = \frac{Z}{\sqrt{Y_2/r}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t(r).$$

Which is useful in hypothesis testing.

- Derivation of the student  $t$ :  $Z = N(0, 1)$ ,  $Y_2 = \chi^2(r)$  with:

$$X_1 = \frac{Z}{\sqrt{Y_2/r}}, \quad X_2 = Y_2.$$

If  $Z$  and  $Y_2$  are independent, we can just multiply their densities together to get the joint distribution and then use our transformation technique to get the joint of  $X_1$  and  $X_2$ . Finally integrate out  $X_2$  to get the distribution of the student  $t$ .

### **F Distribution**

- A random variable  $X$  with density :

$$f(x) = \frac{(r_1/r_2)^{r_1/2} \Gamma[(r_1 + r_2)/2]}{\Gamma(r_1/2) \Gamma(r_2/2)} \frac{x^{(r_1/2)-1}}{(1 + r_1 x/2)^{(r_1+r_2)/2}}, \quad 0 < x < \infty,$$

is distributed  $F(r_1, r_2)$ .  $\mu_x = \frac{r_2}{r_2 - 2}$ . The mean is finite if  $r_2 > 2$ . The variance is finite if  $r_2 > 4$ .

- **Theorem:** If  $Y_1 \sim \chi^2(r_1)$  and  $Y_2 \sim \chi^2(r_2)$  with  $Y_1$  and  $Y_2$  independent. Then,

$$X = \frac{Y_1/r_1}{Y_2/r_2} \sim F(r_1, r_2).$$

- Suppose  $Z \sim N(0, 1)$  and  $Y_2 \sim \chi^2(r_2)$  with  $Z$  and  $Y_2$  are independent. Then:

$$\frac{Z^2}{Y_2/r_2} = \frac{Y_1/1}{Y_2/r_2} \sim F(1, r_2).$$

Recall:

$$\frac{\sqrt{Z^2}}{\sqrt{Y_2/r_2}} = \frac{Z}{\sqrt{Y_2/r_2}} \sim t(r_2).$$

So  $F = t^2$ .

### 16.3 Moment Generating Function Technique

- Suppose we have random variables  $X_1, \dots, X_n$  and new random variables:  $Y_1, \dots, Y_s$  such that:

$$y_i = u_i(x_1, \dots, x_n).$$

- The MGF for the  $y$ 's:

$$\begin{aligned} M_y(t_1, \dots, t_s) &= E \left[ e^{t_1 y_1 + \dots + t_s y_s} \right] \\ &= E \left[ e^{t_1 u_1(x_1, \dots, x_n) + \dots + t_s u_s(x_1, \dots, x_n)} \right] \\ &= \int e^{t_1 u_1(x_1, \dots, x_n) + \dots + t_s u_s(x_1, \dots, x_n)} f(x_1, \dots, x_n) dx_1, \dots, dx_n \end{aligned}$$

If this exists and is recognizable, then you know the probability law of  $Y$  given that there is a 1:1 correspondence between probability laws and MGFs.

- **Theorem:** Suppose  $X_i$ 's are independent RVs with:

$$X_i \sim N(\mu_i, \sigma_i^2).$$

Define:

$$Y = a_0 + \sum_{i=1}^n a_i X_i.$$

Thus:

$$\mu_y = a_0 + \sum_{i=1}^n a_i \mu_i, \quad \sigma_y^2 = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Recall the MGF of  $X$  is  $E[e^{tx}] = e^{t\mu + 1/2\sigma^2 t^2}$ .

- What is the MGF of  $Y$ ?

$$\begin{aligned}
M_y(t) = E[e^{ty}] &= E\left[e^{ta_0+ta_1x_1+\dots+ta_nx_n}\right] \\
&= e^{ta_0} * E[e^{ta_1x_1}] * E[e^{ta_2x_2}] * \dots * E[e^{ta_nx_n}] \\
\text{Let : } & s_i = ta_i \\
&= e^{ta_0} * E[e^{s_1x_1}] * E[e^{s_2x_2}] * \dots * E[e^{s_nx_n}] \\
\text{Note : } & \text{ bunch of normals.} \\
&= e^{ta_0} * e^{s_1\mu_1+1/2\sigma_1^2s_1^2} * e^{s_2\mu_2+1/2\sigma_2^2s_2^2} * \dots * e^{s_n\mu_n+1/2\sigma_n^2s_n^2} \\
\text{Note : } & \text{ back to a's.} \\
&= e^{ta_0} * e^{ta_1\mu_1+1/2\sigma_1^2t^2a_1^2} * e^{ta_2\mu_2+1/2\sigma_2^2t^2a_2^2} * \dots * e^{ta_n\mu_n+1/2\sigma_n^2t^2a_n^2} \\
&= e^{\underbrace{t[a_0 + a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n]}_{\mu_y} + 1/2t^2 \underbrace{[a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2]}_{\sigma_y^2}} \\
&= e^{t\mu_y+1/2t^2\sigma_y^2}
\end{aligned}$$

So  $Y$  is also NORMAL since the MGF of  $Y$  is the same as that of a normal RV. Thus linear functions of normally distributed random variables are NORMAL.

## 17 Lecture 17 - October 28, 2004

### 17.1 Sums and Squares of Normals and $\chi^2$ 's

- If two random variables are independent then the variance of the sum will be greater than or equal to the sum of their variances. If they are not independent, this may not hold. Specifically if  $Y = X_1 + X_2$  and  $X_2 = -X_1$  then  $Y = 0$  and the variance of  $Y$  is 0! Clearly less than the variance of  $X_1$  and  $X_2$ .
- Consider random variables  $X_i \sim iid(\mu, \sigma^2)$ . Then:

$$Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n \sigma^2 = n\sigma^2.$$

$$Var(\frac{1}{n} \sum_{i=1}^n X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

$$Var(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i) = \sigma^2.$$

So note the variance of the sample mean is actually smaller than variance of the individual  $X$ 's.

- **Theorem:** Consider random variables  $X_i \sim \chi^2(r_i)$ , independent. Then:

$$Y = \sum_{i=1}^n X_i \sim \chi^2(r), \quad \text{with } r = \sum_{i=1}^n r_i.$$

Proof: Consider the MGF of  $Y$ :

$$\begin{aligned} M_y(t) = E[e^{tY}] &= E[e^{tx_1 + tx_2 + \dots + tx_n}] \\ &= Ee^{tx_1} * Ee^{tx_2} * \dots * Ee^{tx_n} \\ \text{Note: } \chi^2 \text{ MGF is } E[e^{tx_i}] &= \frac{1}{(1-2t)^{r_i/2}} \\ &= \frac{1}{(1-2t)^{r_1/2}} * \frac{1}{(1-2t)^{r_2/2}} * \dots * \frac{1}{(1-2t)^{r_n/2}} \\ &= \frac{1}{(1-2t)^{r/2}} \end{aligned}$$

So  $Y \sim \chi^2(r)$ .

- **Theorem:** If  $X_i \sim N(\mu_i, \sigma_i^2)$ , independent. Then:

$$Y = \underbrace{\sum_{i=1}^n \underbrace{\left( \underbrace{\frac{X_i - \mu_i}{\sigma_i}}_{N(0,1)} \right)^2}_{\chi^2(1)}}_{\chi^2(n)} \sim \chi^2(n).$$

## 17.2 Multivariate Normal Distribution

- First some notation. Consider a vector of random variables:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

With  $E[X_i] = \mu_i$  and  $cov(X_i, X_j) = \sigma_{ij}$ . Then,

$$E[X] = \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Now consider the following matrix multiplication:

$$E[(x - \mu)(x - \mu)'] = E \left[ \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_n - \mu_n \end{bmatrix} * [x_1 - \mu_1 \dots x_n - \mu_n] \right].$$

So:

$$E[(x - \mu)(x - \mu)'] = E \left[ \begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1) & \dots & (x_1 - \mu_1)(x_n - \mu_n) \\ \vdots & (x_i - \mu_i)(x_j - \mu_j) & \vdots \\ (x_n - \mu_n)(x_1 - \mu_1) & \dots & (x_n - \mu_n)(x_n - \mu_n) \end{bmatrix} \right].$$

Or:

$$E[(x - \mu)(x - \mu)'] = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & \sigma_{ij} & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{bmatrix} = VC(X) = (\sigma_{ij}).$$

- **Definition:** So a vector RV  $X = (X_1, \dots, X_n)'$  with joint density:

$$f(x) = f(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} e^{-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)},$$

with  $\mu = (\mu_1, \dots, \mu_n)'$  and  $\Sigma = (\sigma_{ij})$ , the  $n \times n$  positive definite symmetric variance covariance matrix: we then say that  $X \sim N(\mu, \Sigma)$ .

- Note if  $n = 1$ , we just get back to the univariate normal.
- Proposition.  $f(x) \geq 0$  and  $\int f(x) = 1$ . Not proved.
- Suppose  $Z_1, \dots, Z_n \sim iid N(0, 1)$ . Then (because they are independent) the joint density is:

$$f(z_1, \dots, z_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(1/2)z_i^2}. \quad (1)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-(1/2)\sum z_i^2}. \quad (2)$$

$$= \left(\frac{1}{\sqrt{(2\pi)^n}}\right) e^{-(1/2)Z'Z}. \quad (3)$$

$$(4)$$

Where  $Z = (z_1, \dots, z_n)'$ . So  $Z \sim N(0_{n \times 1}, I_{n \times n})$ . Thus the joint distribution of a bunch of iid normal variables is also normal.

- Note the following linear algebra rules in what comes next:

$$(AB)^{-1} = B^{-1}A^{-1}.$$

$$(A')^{-1} = (A^{-1})'.$$

$$|AB| = |A||B|.$$

$$|A'| = |A|.$$

- So consider the density of  $Z$  again:

$$f(Z) = \frac{1}{\sqrt{(2\pi)^n}} e^{-(1/2)Z'Z}.$$

Since  $\Sigma$  is a positive definite matrix, decompose such that:

$$\Sigma = pp'.$$

Thus,

$$\Sigma^{-1} = p'^{-1}p^{-1}.$$

Now define a new random variable:

$$X = PZ + \mu = \begin{bmatrix} p_{11}z_1 + \dots + p_{1n}z_n + \mu_1 \\ \vdots \\ p_{n1}z_1 + \dots + p_{nn}z_n + \mu_n \end{bmatrix}.$$

What is the distribution of  $X$ ? We'll use the change of variables technique. First solve for  $Z$  in terms of  $X$  because we'll need that soon.

$$p^{-1}X = Z + p^{-1}\mu.$$

$$Z = p^{-1}(X - \mu) = \begin{bmatrix} p^{11}(x_1 - \mu_1) + \cdots + p^{1n}(z_n + \mu_n) \\ \vdots \\ p^{n1}(x_1 - \mu_1) + \cdots + p^{nn}(z_n + \mu_n) \end{bmatrix}.$$

Where  $p^{ij}$  is the  $(i, j)^{th}$  element of  $p^{-1}$ . Note that  $Z' = (X - \mu)'p'^{-1}$ . So since  $Z$  is multivariate normal,

$$g(x) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}(X-\mu)' \underbrace{p'^{-1}p^{-1}}_{\Sigma^{-1}} (X-\mu)} |J|.$$

$$g(x) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)} |J|.$$

Given our definition of  $Z$ , the jacobian is as follows:

$$|J| = \left| \frac{\partial z_i}{\partial x_j} \right| = |p^{-1}| = \frac{1}{|p|} = \frac{1}{\sqrt{|p||p'|}} = \frac{1}{|pp'|^{1/2}} = \frac{1}{|\Sigma|^{1/2}}.$$

Substituting:

$$g(x) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)} \frac{1}{|\Sigma|^{1/2}}.$$

$$g(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)}.$$

And this is the PDF of the multivariate normal. So we have constructed the multivariate normal out of iid univariate  $N(0, 1)$  random variables.

- Proposition. Suppose  $X \sim N(\mu, \Sigma)$ . Then  $X$  has MGF:

$$M(t) = e^{\mu't + (1/2)t'\Sigma t}.$$

Where  $t = (t_1, \dots, t_n)'$ .

- Proposition. If  $X \sim N(\mu, \Sigma)$ . The  $E[X] = \mu$  and  $VC(X) = \Sigma$ . This is easily verified by differentiating the MGF.
- Suppose  $X$  is a RV of dimension  $n \times 1 \sim N(\mu, \Sigma)$ . Define a new variable  $Z$  as follows:

$$Z = AX + c = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n + c_1 \\ \vdots \\ a_{s1}x_1 + \cdots + a_{sn}x_n + c_n \end{bmatrix}.$$

So  $A$  is  $s \times n$ ,  $c$  is  $s \times 1$  and therefore  $Z$  is  $s \times 1$ . So  $Z$  is a linear combination of normal random variables. THEN:

$$Z \sim N(A\mu + c, A\Sigma A').$$

Proof:

$$E[Z] = E[AX + c] = AE[X] + c = A\mu + c.$$

$$VC(Z) = E[(Z - \mu_z)(Z - \mu_z)'].$$

But  $Z - \mu_z = AX + c - A\mu - c = A(X - \mu)$ . So,

$$VC(Z) = E[A(X - \mu)(X - \mu)'A'].$$

$$VC(Z) = AE[(X - \mu)(X - \mu)']A'.$$

$$VC(Z) = A\Sigma A'.$$

QED.

- Proposition. Suppose we have a vector RV,  $X = (X_1, \dots, X_n)'$  and we partition it into:

$$X = \begin{bmatrix} \underbrace{X_1}_{s \times 1} \\ \underbrace{X_2}_{n-s \times 1} \end{bmatrix} \sim N(\mu, \Sigma).$$

With,

$$\mu = \begin{bmatrix} \underbrace{\mu_1}_{s \times 1} \\ \underbrace{\mu_2}_{n-s \times 1} \end{bmatrix}, \text{ and, } \Sigma = \begin{bmatrix} \underbrace{\Sigma_{11}}_{s \times s} & \underbrace{\Sigma_{12}}_{s \times n-s} \\ \underbrace{\Sigma_{21}}_{n-s \times s} & \underbrace{\Sigma_{22}}_{n-s \times n-s} \end{bmatrix}.$$

Then:

$$X_1 \sim N(\mu_1, \Sigma_{11}).$$

- Corollary. If a vector RV  $X \sim N(\mu, \Sigma)$ , then the components of  $X$ :

$$X_i \sim N(\mu_i, \sigma_{ii}).$$

# 18 Lecture 18 - November 2, 2004

## 18.1 More on the Multivariate Normal

- Suppose  $X$  is a random vector:

$$X = \begin{bmatrix} \underbrace{X^1}_{sx1} \\ \underbrace{X^2}_{n-sx1} \end{bmatrix} \sim N(\mu, \Sigma).$$

With,

$$\mu = \begin{bmatrix} \underbrace{\mu^1}_{sx1} \\ \underbrace{\mu^2}_{n-sx1} \end{bmatrix}, \text{ and, } \Sigma = \begin{bmatrix} \underbrace{\Sigma_{11}}_{sx1} & \underbrace{\Sigma_{12}}_{sx1} \\ \underbrace{\Sigma_{21}}_{n-sx1} & \underbrace{\Sigma_{22}}_{n-sx1-s} \end{bmatrix}.$$

Then:

$$X^1 \sim N(\mu^1, \Sigma_{11}).$$

Proof: Consider:  $X^1 = [I_s, 0][X^1, X^2]' = AX$ . So  $X^1$  is a linear combination of normally distributed RVs. Thus  $X^1$  is normal. Or:

$$X^1 \sim N(\underbrace{A\mu}_{\mu^1}, \underbrace{A\Sigma A'}_{\Sigma_{11}}).$$

- Proposition 6. Suppose  $X$  is as above. Then the conditional distribution of  $X^1$  given  $X^2$  is:

$$X^1|X^2 \sim N(\mu^1 + \Sigma_{12}\Sigma_{22}^{-1}(X^2 - \mu^2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Note that if  $\Sigma_{12} = 0$ ,

$$X^1|X^2 \sim N(\mu^1, \Sigma_{11}) \implies \text{Independent}.$$

So for normally distributed random variables,

$$\text{Covariance} = 0 \iff \text{Independence}.$$

BUT ONLY FOR NORMAL RVs.

- Digression. We know the  $\Sigma$  is a positive definite matrix. How can we prove it?  $\Sigma$  is positive semidefinite if:

$$a'\Sigma a \geq 0 \forall a.$$

Suppose  $X$  is a RV with Var/Cov matrix  $\Sigma$ . Consider  $Z = a'X$ . Then:

$$VC(Z) = VC(a'X) = a'VC(X)a = a'\Sigma a \geq 0.$$

So  $\Sigma$ , the Var/Cov matrix, is positive (semi)-definite.

- Proposition 7. Suppose  $Z \sim N(0, \sigma^2 I)$ . This implies that  $Z_i \sim \text{iid } N(0, \sigma^2)$  because of the analysis above. Suppose  $A$  is a symmetric idempotent matrix of rank  $r$ :

$$A = A', \quad AA = A.$$

Then we have:

$$\frac{Z'AZ}{\sigma^2} = \left(\frac{Z}{\sigma}\right)' A \left(\frac{Z}{\sigma}\right) \sim \chi^2(r).$$

Note that  $(Z_i/\sigma) \sim \text{iid } N(0, 1)$ .

PROOF: Decompose  $A$  as follows:

$$CAC' = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Where  $C'C = CC' = I$ , so  $C$  is an orthogonal matrix. Note if you pre-multiply by  $C'$  and post-multiply by  $C$ , we get:

$$(C')CAC'(C) = A = C' \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} C.$$

Now partition the  $C$  matrix into two parts:

$$C = \begin{bmatrix} \underbrace{C_1}_{r \times n} \\ \underbrace{C_2}_{(n-r) \times n} \end{bmatrix}.$$

Thus,

$$I = \begin{bmatrix} I_r & 0 \\ 0 & I_{n-r} \end{bmatrix} = CC' = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} [C_1' \quad C_2'] = \begin{bmatrix} C_1 C_1' & C_1 C_2' \\ C_2 C_1' & C_2 C_2' \end{bmatrix}.$$

Thus:

$$\begin{bmatrix} I_r & 0 \\ 0 & I_{n-r} \end{bmatrix} = \begin{bmatrix} C_1 C_1' & C_1 C_2' \\ C_2 C_1' & C_2 C_2' \end{bmatrix} \implies C_1 C_1' = I_r.$$

Returning to  $A$  above and substituting in the partitioned  $C$ :

$$A = C' \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} C = [C_1' \quad C_2'] \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = C_1' C_1.$$

So back to our quadratic form:

$$Q = \frac{Z'AZ}{\sigma^2} = \frac{Z'C_1' C_1 Z}{\sigma^2} = \underbrace{\frac{Z'C_1'}{\sigma}}_{\eta'} \underbrace{\frac{C_1 Z}{\sigma}}_{\eta_1} = \eta_1' \eta_1.$$

Now recall that:

$$\frac{Z}{\sigma} \sim N(0, 1) \implies \eta_1 = \frac{C_1 Z}{\sigma} \sim N(0, C_1 I C_1') = N(0, I_r).$$

Thus,

$$\eta_1' \eta_1 \sim \chi^2(r),$$

Because we have the sum of  $r$  squared normals which is a  $\chi^2$  variable where the degrees of freedom add up. QED.

- Proposition 8. Consider the same setup as in proposition 7 and consider two forms:

$$\underbrace{Q = \frac{Z' A Z}{\sigma^2}}_{\text{Quadratic Form}}, \quad \underbrace{X = \frac{B Z}{\sigma}}_{\text{Linear Form}}.$$

Assume  $BA = 0$ . Then:  $Q$  and  $X$  are INDEPENDENT.

Proof: As before, denote:

$$\eta_1 = \frac{C_1 Z}{\sigma} \sim N(0, I_r),$$

and denote:

$$\eta_2 = \frac{C_2 Z}{\sigma} \sim N(0, I_{n-r}).$$

Thus,

$$E[\eta_1 \eta_2'] = E\left[\frac{C_1 Z}{\sigma} \frac{Z' C_2'}{\sigma}\right] = C_1 E\left[\underbrace{\frac{Z Z'}{\sigma \sigma}}_I\right] C_2' = C_1 C_2' = 0.$$

Thus  $\eta_1$  and  $\eta_2$  are independent. Recall our quadratic form depends only on  $\eta_1$ .

$$Q = \frac{Z' A Z}{\sigma^2} = \eta_1' \eta_1.$$

Now consider our linear form:

$$X = \frac{B Z}{\sigma} = B \underbrace{C' C}_I \frac{Z}{\sigma} = B C' \eta.$$

Now let:

$$B C' = D = \begin{bmatrix} \underbrace{D_1}_{m \times r} & \underbrace{D_2}_{m \times (n-r)} \end{bmatrix}.$$

When  $D_2 = 0$ ,

$$[D_1, 0] = D \underbrace{\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}}_{C A C'} = D C A C' = B \underbrace{C' C}_I A C' = \underbrace{B A}_0 C' = 0.$$

Thus  $D_1 = 0$ . So our linear form becomes:

$$B\frac{Z}{\sigma} = BC'C\frac{Z}{\sigma} = D\frac{CZ}{\sigma} = D\eta = \begin{bmatrix} 0 & D_2 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = D_2\eta_2.$$

So since the linear form only depends on  $\eta_2$  and the quadratic form only depends on  $\eta_1$ , the linear and quadratic forms are INDEPENDENT! QED.

## 19 Lecture 19: November 9, 2004

- Proposition 9. Suppose  $Z \sim N(0, \sigma^2 I)$  and suppose  $A$  and  $B$  are  $n \times n$  symmetric matrices such that:

$$BA = 0.$$

Define the following two quadratic forms:

$$Q_1 = Z'AZ, \text{ and } Q_2 = Z'BZ.$$

Then  $Q_1$  and  $Q_2$  are independent.

- Note that if you have a quadratic form,  $Z'AZ$ , and  $A$  is NOT symmetric, then note that  $Z'AZ = Z'A'Z$  because this is just a scalar, so:

$$Z'AZ = \frac{1}{2}(Z'AZ + Z'AZ) = Z' \underbrace{\left[ \frac{1}{2}(A + A) \right]}_B Z,$$

where  $B$  is now symmetric.

### 19.1 Distributions of the Sample Mean and Sample Variance

- **Theorem:** Suppose  $X_1, \dots, X_n$  is a random sample with  $X_i \sim iid N(0, \sigma^2)$ . Thus,

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

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

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

We will show that the sample mean and variance are independent.

- Let  $e = [1, \dots, 1]'$ , an  $n \times 1$  vector of ones. Thus,

$$\bar{X} = \frac{1}{n} e'X.$$

- Deviations from the mean are then  $X - e\bar{X}$  and substituting in the sample mean:

$$X - e\bar{X} = X - e \frac{1}{n} e'X = \underbrace{\left( I - \frac{ee'}{n} \right)}_A X.$$

- Note that  $A$  is symmetric so  $A = A'$ . Also note:

$$Ae = \left(I - \frac{ee'}{n}\right)e = e - \frac{ee'e}{n} = e - \frac{en}{n} = 0,$$

because  $e'e = n$ . So  $A$  and  $e$  are orthogonal. Also,

$$\begin{aligned} AA &= \left(I - \frac{ee'}{n}\right)\left(I - \frac{ee'}{n}\right) = I - \frac{ee'}{n} - \frac{ee'}{n} + \frac{ee'ee'}{n^2} = \\ &= I - \frac{ee'}{n} - \frac{ee'}{n} + \frac{ene'}{n^2} = I - \frac{ee'}{n} - \frac{ee'}{n} + \frac{ee'}{n} = I - \frac{ee'}{n} = A. \end{aligned}$$

So  $A$  is idempotent.

- Let  $Z = X - e\bar{X} = \left(I - \frac{ee'}{n}\right)X = AX$ . And thus:

$$nS^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = Z'Z.$$

Substitute for  $Z$ ,

$$nS^2 = X'A'AX = X'AX.$$

Divide both sides by  $\sigma^2$ ,

$$\frac{nS^2}{\sigma^2} = \frac{X'AX}{\sigma^2} \sim \chi^2(\text{rank}(A)).$$

Recall:

$$\text{rank}(A) = \text{tr}(A) = \text{tr}\left(I_n - \frac{ee'}{n}\right) = \text{tr}(I_n) - \frac{1}{n}\text{tr}(ee') = n - \frac{1}{n}\text{tr}\left(\underbrace{e'e}_n\right) = n - 1.$$

So,

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n - 1).$$

- Recall the sample mean:

$$\bar{X} = \frac{1}{n}e'X = BX \sim N\left(0, \frac{\sigma^2}{n}\right).$$

- Note that:

$$BA = \frac{1}{n}e' * \left(I - \frac{ee'}{n}\right) = \frac{1}{n}e' - \frac{e'ee'}{n^2} = \frac{e'}{n} - \frac{e'}{n} = 0.$$

So, since  $\bar{X}$  is a linear form and  $nS^2$  is a quadratic form with  $A$  as an idempotent, symmetric matrix and  $BA = 0$ , then  $\bar{X}$  and  $nS^2$  are independent by a previous proposition.

- Suppose  $X_i \sim iid (\mu, \sigma^2)$ . Then,

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu),$$

has:

$$E[Z_n] = 0.$$

$$Var(Z_n) = \sigma^2.$$

By the central limit theorem, it can be shown that the distribution of  $Z_n$  converges to a normal in the limit. Thus,

$$Z_n \simeq N(0, \sigma^2).$$

Thus,

$$\begin{aligned} Z_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i - \frac{1}{\sqrt{n}} n * \mu \\ &= \frac{1}{\sqrt{n}} \bar{X} * n - \frac{n}{\sqrt{n}} \mu \\ &= \frac{n}{\sqrt{n}} (\bar{X} - \mu) \\ &= \frac{n\sqrt{n}}{n} (\bar{X} - \mu) \\ &= \sqrt{n} (\bar{X} - \mu), \end{aligned}$$

then,

$$\bar{X} = \mu + \frac{1}{\sqrt{n}} Z_n.$$

Which implies,

$$\bar{X} \simeq N\left(\mu, \frac{1}{\sqrt{n}^2} Var(Z_n)\right) = N\left(\mu, \frac{\sigma^2}{n}\right).$$

- **Theorem:** Suppose  $X_1, \dots, X_n$  is a random sample with  $X_i \sim iid N(0, \sigma^2)$ . Then:

$$\sqrt{n-1} \frac{\bar{X} - \mu}{S} \sim t(n-1).$$

Proof: Start with

$$\sqrt{n-1} \frac{\bar{X} - \mu}{S}$$

Divide top and bottom by  $\sigma$

$$\sqrt{n-1} \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{S}}}{\sigma}$$

Square and root bottom:

$$\sqrt{n-1} \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{S}}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

Multiply top and bottom by  $\sqrt{n}$ :

$$\sqrt{n-1} \frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\frac{\sigma}{S}}}{\sqrt{\frac{nS^2}{\sigma^2}}}$$

Move the root  $n-1$  to the bottom of the bottom:

$$\frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\frac{\sigma}{S}}}{\sqrt{\frac{nS^2}{\sigma^2} / (n-1)}}$$

Move the root  $n$  to the bottom of the bottom of the top:

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma/\sqrt{n}}{S}}}{\sqrt{\frac{nS^2}{\sigma^2} / (n-1)}}$$

Note the top is distributed  $N(0, 1)$  and the bottom is a  $\chi^2(n-1)$  divided by its degrees of freedom  $(n-1)$ . Thus,

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma/\sqrt{n}}{S}}}{\sqrt{\frac{nS^2}{\sigma^2} / (n-1)}} \sim t(n-1).$$

QED.

## 19.2 Application to Regression Analysis

- Suppose we have a regression model:

$$y_i = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_k z_{ik} + u_i, \quad u_i \sim iid N(0, \sigma^2).$$

When we say  $X_i \sim iid N(\mu, \sigma^2)$ , we could also write the special linear regression model:

$$X_i = \mu + u_i, \quad u_i \sim iid N(0, \sigma^2).$$

The OLS estimator for  $\alpha$  is generally  $(Z'Z)^{-1}Z'y$  and in our case:

$$\hat{\mu} = (e'e)^{-1}e'X = \frac{1}{n} \sum X_i = \bar{X}.$$

So the OLS estimator is equal to the sample mean.

- Now the variance of the OLS estimator is usually:

$$\sigma_{\hat{\alpha}}^2 = (Z'Z)^{-1}\sigma^2.$$

Which in our case reduces to:

$$\sigma_{\hat{\mu}}^2 = (e'e)^{-1}\sigma^2 = \frac{\sigma^2}{n}.$$

To estimate  $\sigma^2$ , we generally use:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n \hat{u}_i^2.$$

But  $\hat{u}_i = X_i - \hat{\mu} = X_i - \bar{X}$ ,

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = S_1^2 = \frac{n}{n-1} S^2.$$

So the estimated variance of the estimated coefficient is:

$$\hat{\sigma}_{\hat{\mu}}^2 = (e'e)^{-1}\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}.$$

- Thus a usual t-test gives us the following test statistic:

$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}}.$$

Divide top and bottom by  $\sigma_{\hat{\mu}}$ :

$$T = \frac{(\bar{X} - \mu)/\sigma_{\hat{\mu}}}{\hat{\sigma}_{\hat{\mu}}/\sigma_{\hat{\mu}}}.$$

Note  $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$ .

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\hat{\sigma}_{\hat{\mu}}/\sigma_{\hat{\mu}}}.$$

Root and Square the bottom:

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\hat{\sigma}_{\hat{\mu}}^2/\sigma_{\hat{\mu}}^2}}.$$

Note:  $\hat{\sigma}_{\hat{\mu}}^2 = \frac{\hat{\sigma}^2}{n}$  and  $\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}$ , so:

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\frac{\hat{\sigma}^2/n}{\sigma^2/n}}}.$$

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}.$$

Recall  $\hat{\sigma}^2 = \frac{n}{n-1}S^2$ , so:

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\frac{nS^2/(n-1)}{\sigma^2}}}.$$

Or:

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\frac{nS^2}{\sigma^2}/(n-1)}}.$$

Cancelling:

$$T = \sqrt{n-1} \frac{\bar{X} - \mu}{S} \sim t(n-1).$$

Good!

## 20 Lecture 20: November 11, 2004

### 20.1 Review of CDF Technique

- Consider a 2-dimensional random variable  $X = [X_1, X_2]$ . Suppose  $Y = G(X)$ . Then:

$$F_Y(y) = Pr(Y \leq y) = Pr(g(X) \leq y) = \int \int_A f(x_1, x_2) dx_1 dx_2,$$

where  $A = \{(x_1, x_2) : g(x_1, x_2) \leq y\}$ .

### 20.2 Point Estimation

- Suppose  $Y_1, \dots, Y_n$  are iid  $N(\mu, 1)$ . What is the mean of the  $Y$ 's? We might try the sample mean:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = u(Y_1, \dots, Y_n).$$

- Suppose  $Y_i = a + bX_i + u_i$  with  $u_i \sim$  iid  $N(0, 1)$ . Then:

$$Y_i \sim N(a + bX_i, 1).$$

What are  $a$  and  $b$ ? We could estimate via OLS:

$$\hat{b} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

- But in both these cases we just are choosing an estimate of the parameter. We could think of the  $Y_i$ 's to have joint density:

$$F(y_1, \dots, y_n; \theta),$$

where  $\theta = \mu$  in the first case and  $\theta = (a, b)'$  in the second. So we have:

$$\theta \in \Theta = \mathfrak{R},$$

or

$$\theta \in \Theta = \mathfrak{R}x\mathfrak{R}.$$

- Asking what is  $\theta$ ? is equivalent to asking what distribution really generated the  $Y$ 's. We may know the class of distributions that the  $Y$ 's come from, in which case we have a problem of parametric estimation. In this case the parameter vector,  $\theta$ , is finite. If  $Y_i = g(X_i, u_i)$  but we don't know anything about the distribution, then we have a problem of non-parametric estimation. In this case the parameter vector,  $\theta$ , is of infinite dimension. We will focus on parametric estimation. Good.

- There are two principal ways to go about estimating  $\theta$ , Maximum Likelihood Estimation and the Method of Moments Technique.

### 20.3 Maximum Likelihood Estimation (MLE)

- Example 1: Consider a random sample  $X_1, \dots, X_n$  such that  $X_i \sim \text{Bernoulli}(\theta)$ . So:

$$f(x) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1, \quad 0 \leq \theta \leq 1.$$

- Example 2: Consider a random sample  $X_1, \dots, X_n$  such that  $X_i \sim N(\mu, 1)$ .
- Consider example 2 and suppose  $n = 1$ . In this case we have one observation and suppose we can choose from a manifold of distributions each with different  $\mu$ . You will clearly want to choose the one that maximizes the density at the point  $X_1$ . In other words, we would like to choose the distribution such that:

$$\hat{\mu} = \arg \max_{\mu} f(X_1, \mu).$$

This is the maximum likelihood estimator.

- Note that when maximizing a function's density, maximizing a monotonic transformation will also yield the same result. If we have a function  $g(x)$  to maximize and suppose  $h(y)$  is monotonic, then:

$$\text{Max}_x h(g(x)) \Rightarrow \underbrace{\frac{\partial h(g(x))}{\partial g}}_{>0} * \underbrace{\frac{\partial g(x)}{\partial x}}_{=0} = 0.$$

So maximizing the log of a density is going to give us the same solution.

- Back to example 1. Consider the joint density (note iid):

$$\begin{aligned} f(x_1, \dots, x_n; \theta) &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= \theta^{x_1} (1 - \theta)^{1-x_1} * \theta^{x_2} (1 - \theta)^{1-x_2} * \dots * \theta^{x_n} (1 - \theta)^{1-x_n} \\ &= \theta^{\sum x_i} (1 - \theta)^{\sum 1-x_i} \\ &= \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \end{aligned}$$

Here we assume the  $X$ 's are random variables while the  $\theta$  is fixed. Next we write the likelihood function where we now assume the  $X$ 's are realizations and we don't know

$\theta$ :

$$\begin{aligned}
 L(\theta; x_1, \dots, x_n) &= \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \\
 \text{Log } L &= \sum x_i \log(\theta) + (n - \sum x_i) \log(1 - \theta) \\
 \frac{\partial \text{Log } L}{\partial \theta} &= \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta} \\
 0 &= \frac{\sum x_i}{\hat{\theta}} - \frac{n - \sum x_i}{1 - \hat{\theta}} \\
 \frac{\sum x_i}{\hat{\theta}} &= \frac{n - \sum x_i}{1 - \hat{\theta}} \\
 \sum x_i &= n\hat{\theta} \\
 \hat{\theta} &= \frac{1}{n} \sum x_i \implies \text{The Sample Mean!}
 \end{aligned}$$

- So  $\hat{\theta}(\underbrace{x_1, \dots, x_n}_{\text{Realizations}}) = \frac{1}{n} \sum_{i=1}^n x_i$  is our ESTIMATE – A number with Variance ZERO.
- So  $\hat{\theta}(\underbrace{X_1, \dots, X_n}_{\text{Random Variables}})$  is the ESTIMATOR – itself a RANDOM VARIABLE.
- If we don't necessarily know the original distribution but we invoke one, the estimator is called the Quasi-Maximum Likelihood Estimator.
- Example 2. Joint Density:

$$\begin{aligned}
 f(x_1, \dots, x_n; \mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-0.5*(x_i - \mu)^2} \\
 &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^n e^{-0.5*(x_i - \mu)^2} \\
 &= (2\pi)^{-n/2} \prod_{i=1}^n e^{-0.5*(x_i - \mu)^2} \\
 &= (2\pi)^{-n/2} e^{-0.5*\sum (x_i - \mu)^2}
 \end{aligned}$$

Next we write the likelihood function:

$$\begin{aligned}L(\mu; x_1, \dots, x_n) &= (2\pi)^{-n/2} e^{-0.5 \sum (x_i - \mu)^2} \\ \text{Log } L &= (-n/2) \log(2\pi) - 0.5 \sum (x_i - \mu)^2 \\ \frac{\partial \text{Log } L}{\partial \mu} &= \sum (x_i - \mu) \\ 0 &= \sum (x_i - \hat{\mu}) \\ 0 &= \sum (x_i) - n\hat{\mu} \\ \hat{\mu} &= \frac{1}{n} \sum x_i \implies \text{The Sample Mean!}\end{aligned}$$

- So what are good properties of MLEs? Unbiased and minimum variance. More next week.

## 21 Lecture 21: November 16, 2004

### 21.1 Maximum Likelihood Estimation

- Suppose we have a joint PDF,  $f(x_1, \dots, x_n; \theta)$  yielding likelihood function:

$$L(\theta; x_1, \dots, x_n).$$

Suppose,

$$\hat{\theta}_{mx1} = u(x_1, \dots, x_n) = \begin{bmatrix} u_1(x_1, \dots, x_n) \\ \vdots \\ u_m(x_1, \dots, x_n) \end{bmatrix}.$$

So in the end, we have:

$$L[\underbrace{u(x_1, \dots, x_n)}_{\text{Maximizer}}; x_1, \dots, x_n] \geq L(\theta; x_1, \dots, x_n).$$

The maximizer,  $\hat{\theta}$  is found by maximizing  $L$  with respect to the  $\theta$ 's. The MLE is not necessarily unique and the likelihood function need not be differentiable for the MLE to exist.

- So we have:

$$\hat{\theta} = u(x_1, \dots, x_n) : \text{the ML ESTIMATE.}$$

$$\hat{\theta} = u(X_1, \dots, X_n) : \text{the ML ESTIMATOR.}$$

So the estimate is a realization of the estimator.

- Note that ML estimators have good large sample properties. They are asymptotically efficient. More on this later.
- Example.  $X_1, \dots, X_n$  RVs with  $X_i \sim \text{iid } N(\theta_1, \theta_2)$ . The log-likelihood function is therefore:

$$\ln L(\theta_1, \theta_2) = -\frac{\sum (x_i - \theta_1)^2}{2\theta_2} - n \frac{\ln(2\pi\theta_2)}{2}.$$

FOCs:

$$\frac{\partial \ln L}{\partial \theta_1} \Rightarrow \frac{\sum (x_i - \theta_1)}{\theta_2} = 0.$$

$$\frac{\partial \ln L}{\partial \theta_2} \Rightarrow \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2} = 0.$$

Solving yields ESTIMATES:

$$\hat{\theta}_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\theta}_2 = S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

And ESTIMATORS:

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

So  $E[\hat{\theta}_1] = \theta_1$ , which means that the ML estimator for the mean is unbiased. However,

$$\begin{aligned} E[\hat{\theta}_2] &= E\left[\frac{n-1}{n} \frac{1}{n-1} \sum (X_i - \bar{X})^2\right] \\ &= \frac{n-1}{n} E\left[\frac{1}{n-1} \sum (X_i - \bar{X})^2\right] \\ &= \frac{n-1}{n} \theta_2 \implies \text{Biased!!} \end{aligned}$$

However, we will see that although the ML estimator for the variance is biased, a SMALL sample property, it is consistent, a LARGE sample property.

## 21.2 Unbiasedness and Consistency

- **Definition:** Suppose we have RVs,  $X_1, \dots, X_n$  with joint CDF:

$$F(x_1, \dots, x_n; \theta), \quad \theta \in \Theta.$$

$\theta$  is some statistic which we estimate with:

$$\hat{\theta} = u(X_1, \dots, X_n).$$

Then if

$$E_{\theta_0}[\hat{\theta}] = \theta_0 \quad \forall \theta_0 \in \Theta,$$

$\hat{\theta}$  is UNBIASED for  $\theta_0$ . Note the expectation is with respect to true (unknown) parameter.

- Example.  $X_i \sim \text{iid } N(\theta, 1)$ . Then  $\hat{\theta} = \frac{1}{n} \sum X_i$ .  $E[\hat{\theta}] = \frac{1}{n} \sum E[X_i] = \theta \quad \forall \theta \in \Theta$ . So here  $\hat{\theta}$  is an unbiased estimator for the sample mean. Now consider  $\tilde{\theta} = 5$ , so  $E[\tilde{\theta}] = 5$ . Thus  $\tilde{\theta}$  is unbiased, ie  $E[\tilde{\theta}] = \theta$  if  $\theta = 5$ . This is not a strict interpretation of the definition since for an estimator to be unbiased, it must be FOR ALL  $\theta \in \Theta$ . So we would not say that  $\tilde{\theta}$  was unbiased.
- Note that unbiased estimators need not be unique.
- **Definition:** Consistency. An estimator  $\hat{\theta}_n = u(X_1, \dots, X_n)$  is (weakly) consistent if:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0, \quad \forall \epsilon > 0.$$

Where we index the estimator with  $n$  corresponding to the sample size.

- Sufficiency condition for consistency. Consider the following bounds on the above probability:

$$0 \leq P(|\hat{\theta}_n - \theta| \geq \epsilon)$$

$$0 \leq P((\hat{\theta}_n - \theta)^2 \geq \epsilon^2)$$

By Chebychev:

$$0 \leq P((\hat{\theta}_n - \theta)^2 \geq \epsilon^2) \leq \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}$$

$$\lim_{n \rightarrow \infty} 0 \leq \lim_{n \rightarrow \infty} P((\hat{\theta}_n - \theta)^2 \geq \epsilon^2) \leq \lim_{n \rightarrow \infty} \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}$$

$$0 \leq \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \lim_{n \rightarrow \infty} \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}$$

Now consider the term in the numerator of the last term:

$$\begin{aligned} E[(\hat{\theta}_n - \theta)^2] &= E[(\hat{\theta}_n - E[\hat{\theta}_n] + E[\hat{\theta}_n] - \theta)^2] \\ &= E[(\hat{\theta}_n - E[\hat{\theta}_n])^2 + (E[\hat{\theta}_n] - \theta)^2 + 2(\hat{\theta}_n - E[\hat{\theta}_n]) * (E[\hat{\theta}_n] - \theta)] \\ &= E[(\hat{\theta}_n - E[\hat{\theta}_n])^2] + E[(E[\hat{\theta}_n] - \theta)^2] + 2E[(\hat{\theta}_n - E[\hat{\theta}_n]) * (E[\hat{\theta}_n] - \theta)] \\ &= E[(\hat{\theta}_n - E[\hat{\theta}_n])^2] + E[(E[\hat{\theta}_n] - \theta)^2] + 2(E[\hat{\theta}_n] - \theta)E[(\hat{\theta}_n - E[\hat{\theta}_n])] \\ &= E[(\hat{\theta}_n - E[\hat{\theta}_n])^2] + E[(E[\hat{\theta}_n] - \theta)^2] + 2(E[\hat{\theta}_n] - \theta) \underbrace{E[(\hat{\theta}_n - E[\hat{\theta}_n])]}_0 \\ &= E[(\hat{\theta}_n - E[\hat{\theta}_n])^2] + E[(E[\hat{\theta}_n] - \theta)^2] \\ &= \text{Var}(\hat{\theta}_n) + [\text{Bias}(\hat{\theta}_n)]^2 \\ &= \text{Mean Square Error} \end{aligned}$$

Thus if the MSE is 0, or:

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0, \quad \lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta,$$

then (from above):

$$0 \leq \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \underbrace{\lim_{n \rightarrow \infty} \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}}_0.$$

$$0 \leq \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) \leq 0.$$

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0.$$

Consistent!

- So a sufficient set of conditions for consistency is that the variance of the estimator must go to 0 and the expected value of the estimator must “converge in probability” to the true parameter. Thus an unbiased estimator (the second condition) does not necessarily imply consistency.

### 21.3 Method of Moments Technique

- Example. Random variables  $X_1, \dots, X_n \sim \Gamma(\alpha, \beta)$  so  $\theta_1 = \alpha$  and  $\theta_2 = \beta$  are the two parameters we would like to estimate. For a gamma distribution, we have the first two moments defined by:

$$\mu = \alpha\beta = \theta_1\theta_2,$$

and,

$$\sigma^2 = \alpha\beta^2 = \theta_1\theta_2^2.$$

Thus we might take the data on the  $X$ 's and estimate  $\hat{\mu}$  and  $\hat{\sigma}^2$ . Then equate as follows:

$$\begin{aligned}\hat{\mu} &= \hat{\theta}_1\hat{\theta}_2, \\ \hat{\sigma}^2 &= \hat{\theta}_1\hat{\theta}_2^2,\end{aligned}$$

Two equations and two unknowns yields estimators:

$$\begin{aligned}\hat{\theta}_1 &= \frac{\hat{\mu}^2}{\hat{\sigma}^2} = \frac{\bar{x}^2}{S^2}, \\ \hat{\theta}_2 &= \frac{\hat{\sigma}^2}{\hat{\mu}} = \frac{S^2}{\bar{x}}.\end{aligned}$$

And these are our estimators based on the first two moments of the distribution.

## 22 Lecture 22: November 18, 2004

### 22.1 More on the Method of Moments

- Suppose we have RV's,  $X_1, \dots, X_n$ , with joint distribution,  $F(X, \theta_1, \dots, \theta_k)$ . Then define the population moments:

$$E[X_i^1] = g_1(\theta_1, \dots, \theta_k) = \int x^1 f(x, \theta_1, \dots, \theta_k) dx,$$

$$E[X_i^2] = g_2(\theta_1, \dots, \theta_k) = \int x^2 f(x, \theta_1, \dots, \theta_k) dx,$$

$\vdots$

$$E[X_i^k] = g_k(\theta_1, \dots, \theta_k) = \int x^k f(x, \theta_1, \dots, \theta_k) dx,$$

Which defines:

$$\theta_i = h_i(\mu_1, \dots, \mu_k), \quad i = 1 \dots k.$$

- Then we estimate the sample moments:

$$\hat{\mu}_m = \frac{1}{n} \sum x_i^m = g_m(\hat{\theta}_1, \dots, \hat{\theta}_k).$$

Which defines:

$$\hat{\theta}_i = h_i(\hat{\mu}_1, \dots, \hat{\mu}_k), \quad i = 1 \dots k.$$

- So suppose we have 3 parameters to estimate and we start with the mean and variance:

$$\mu_1 = E[X_i] = g_1(\theta_1, \theta_2, \theta_3).$$

$$\sigma^2 = E[(X_i - \mu)^2] = g_2(\theta_1, \theta_2, \theta_3).$$

We would clearly need (at least) one additional moment to estimate the 3 parameters. So how about:

$$\mu_3 = E[(X_i - \mu)^3] = g_3(\theta_1, \theta_2, \theta_3).$$

We could then estimate the three moments with:

$$\hat{\mu}_1 = \frac{1}{n} \sum X_i = g_1(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3).$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = g_2(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3).$$

$$\hat{\mu}_3 = \frac{1}{n} \sum (X_i - \bar{X})^3 = g_3(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3).$$

- The key is to have at least  $k$  independent equations to estimate the  $k$  parameters. The choice of which to choose is more subtle and indeed, you could estimate them all (all

finite moments at least) and then invoke the Generalized Method of Moments (GMM) to find the estimators.

- Example. Suppose we have the following regression model:

$$y_i = a + bx_i + u_i, \quad u_i \sim (0, \sigma^2).$$

Assume the  $x$ 's are non-stochastic. Then:

$$E[x_i u_i] = 0.$$

And:

$$E[u_i] = 0.$$

Thus, to estimate our parameters ( $a$  and  $b$  in this case), try the following sample moments:

$$\begin{aligned} \frac{1}{n} \sum u_i &= \frac{1}{n} \sum y_i - a - bx_i = \frac{1}{n} \sum y_i - a - b \frac{1}{n} \sum x_i. \\ \frac{1}{n} \sum x_i u_i &= \frac{1}{n} \sum x_i (y_i - a - bx_i) = \frac{1}{n} \sum x_i y_i - a \frac{1}{n} \sum x_i - b \frac{1}{n} \sum x_i^2. \end{aligned}$$

If we set these equal to zero, as we know they hold in expectation, we cannot solve for  $a$  and  $b$ , but instead must replace them with their estimates:

$$\begin{aligned} 0 &= \frac{1}{n} \sum y_i - \hat{a} - \hat{b} \frac{1}{n} \sum x_i. \\ 0 &= \frac{1}{n} \sum x_i y_i - \hat{a} \frac{1}{n} \sum x_i - \hat{b} \frac{1}{n} \sum x_i^2. \end{aligned}$$

But these are exactly the normal equations for the OLS estimators!

- Thus the ML and MM estimators correspond exactly to the OLS estimators.

## 22.2 Quality of Estimators

- **Definition:** Suppose  $X_1, \dots, X_n$  are random variables with distribution  $F(X; \theta)$ . Suppose we have the following estimator of  $\theta$ :

$$\hat{\theta} = u_1(X_1, \dots, X_n).$$

Define the Mean Square Error:

$$MSE(\theta) = E_\theta(\hat{\theta} - \theta)^2 = \int \cdots \int [u(x_1, \dots, x_n) - \theta]^2 f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n.$$

See G-22.1 for a graph of the  $MSE$  for two estimators. If the  $MSE$  is above the  $MSE$  of another estimator for all values of the true parameter, then we can clearly say that one is better than the other. However, usually (as in the second graph) this is NOT the case.

- **Definition:** Unbiased Minimum Variance Estimator. Suppose we again have our estimator:

$$\hat{\theta} = u_1(X_1, \dots, X_n).$$

Now make the following two restrictions:

- (1)  $E[\hat{\theta}] = \theta$ .
- (2)  $Var(\hat{\theta}) \leq Var(\tilde{\theta})$  for any  $\tilde{\theta}$  where  $E[\tilde{\theta}] = \theta$ .

Note that if  $E[\tilde{\theta}] = \theta$ , then  $MSE_{\tilde{\theta}} = Var(\tilde{\theta})$  since the  $MSE$  is generally equal to the variance of the estimator plus the square of the bias. Essentially what we are doing here is limiting ourselves to unbiased estimators, and the finding the one that has the smallest variance ( $MSE$ ).

- In econometrics we often look at  $E[(\hat{\theta} - \theta)^2]$  as our loss function instead of say  $E|\hat{\theta} - \theta|$  or something which may weight the positive deviations more than the negative or vice versa. Ideally, you would look at every problem individually and determine the best measure of estimator quality based on the particulars of the problem. When this is not feasible, the  $MSE$  is used most often.

## 22.3 Reliability of Estimators - Confidence Intervals

- Consider the sample mean of the random variables  $X_1, \dots, X_n$  with  $F(X_1, \dots, X_n; \theta)$ . We look to find values  $Y_1$  and  $Y_2$  such that:

$$Pr(Y_1 < \theta < Y_2) = 0.95,$$

ie, form a 95% confidence interval around  $\theta$ . In general:

$$Y_1 = u_1(X_1, \dots, X_n),$$

$$Y_2 = u_2(X_1, \dots, X_n).$$

Next time we'll figure out how to find the values of  $Y_1$  and  $Y_2$  when  $\theta$  is the population mean.

## 23 Lecture 23: November 23, 2004

### 23.1 More on Confidence Intervals

- Suppose we have random variables  $X_1, \dots, X_n$  with distribution,  $F(X_1, \dots, X_n; \theta)$ , and we estimate two values:

$$Y_1 = u_1(x_1, \dots, x_n).$$

$$Y_2 = u_2(x_1, \dots, x_n).$$

Then,

$$Pr(Y_1 \leq \theta \leq Y_2) = \gamma \Rightarrow (Y_1, Y_2) \text{ is a } \gamma\% \text{ Confidence Interval for } \theta.$$

- Suppose  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ . If  $\sigma^2$  is known, note:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Let the CDF of the normal(0, 1) be defined as  $N(z)$  such that:

$$N(z_\gamma) = Pr(Z \leq z_\gamma) = \gamma.$$

Where  $z_\gamma$  is the  $\gamma$  fractile of the normal(0, 1).

- If  $\gamma = 0.95$ , we split this area between the left and right tails to get  $(\gamma + 1)/2 = 0.975$ . See G-23.1. Thus,

$$\begin{aligned} Pr(-z_{(\gamma+1)/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{(\gamma+1)/2}) &= \gamma \\ Pr(-z_{(\gamma+1)/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \leq z_{(\gamma+1)/2}) &= \gamma \\ Pr(-\bar{X} - z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}) &= \gamma \\ Pr(\bar{X} + z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}) &= \gamma \\ Pr(\underbrace{\bar{X} - z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}}_{Y_1} \leq \mu \leq \underbrace{\bar{X} + z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}}_{Y_2}) &= \gamma \end{aligned}$$

So our  $\gamma\%$  confidence interval is:

$$\bar{X} \pm z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}.$$

- Now suppose  $\sigma^2$  is unknown. Then:

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}} \sim t(n-1).$$

- And similarly,

$$\begin{aligned} Pr(-t_{(\gamma+1)/2, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n-1}} \leq t_{(\gamma+1)/2, n-1}) &= \gamma \\ Pr(\underbrace{\bar{X} - t_{(\gamma+1)/2, n-1} \frac{S}{\sqrt{n-1}}}_{Y_1} \leq \mu \leq \underbrace{\bar{X} + t_{(\gamma+1)/2, n-1} \frac{S}{\sqrt{n-1}}}_{Y_2}) &= \gamma \end{aligned}$$

So our  $\gamma\%$  confidence interval for  $\mu$  is:

$$\bar{X} \pm t_{(\gamma+1)/2, n-1} \frac{S}{\sqrt{n-1}}.$$

- So since the  $t$  has heavier tails than the  $Z$ , the confidence intervals will be larger for a  $t$ . This is intuitive since now we don't know  $\sigma^2$  so we should have larger intervals.

## 23.2 Testing Statistical Hypothesis

### Examples and Definition

- Suppose  $X \sim N(\theta, 100)$  and we want to test:

$$H_0 : \theta \leq 75.$$

*vs*

$$H_1 : \theta > 75.$$

- Let  $\Omega = \{(x_1, \dots, x_n) : -\infty < x_i < \infty\} = \mathfrak{R}^n$ , or the set of ALL possible realizations. Partition the set as follows:

$$\Omega = C \cup C^*, \text{ with } C \cap C^* = \emptyset.$$

Where  $C$  is called the “Critical Region” where if  $(x_1, \dots, x_n) \in C$ , then we reject  $H_0$ . Note that this set  $C$  completely defines the test.

- Test 1. Assume  $n = 25$  so our critical region could be defined as:

$$C = \{(x_1, \dots, x_n) : X_1 + X_2 + \dots + X_n > 25 * 75\} = \{(x_1, \dots, x_n) : \bar{X} > 75\}.$$

Note  $\bar{X} \sim N(\theta, \sigma^2/n) = N(\theta, 4)$ . So define the probability of a concluding a rejection of the null as:

$$\begin{aligned}
 K_1(\theta) &= Pr((x_1, \dots, x_n) \in C) \\
 &= Pr(\bar{X} \geq 75) \\
 &= Pr\left(\frac{\bar{X} - \theta}{2} \geq \frac{75 - \theta}{2}\right) \\
 &= Pr\left(Z \geq \frac{75 - \theta}{2}\right) \\
 &= 1 - Pr\left(Z \leq \frac{75 - \theta}{2}\right) \\
 &= 1 - N\left(\frac{75 - \theta}{2}\right)
 \end{aligned}$$

Note that:

$$K_1(73) = 0.158, \quad K_1(75) = 0.5, \quad K_1(79) = 0.841.$$

If the true mean is 75, and we use  $C$  defined above, we will reject the null (incorrectly!) half the time. This  $K_1(\theta)$  function is called the Power function. If the true mean is 79 and we use the  $C$  defined above, we will correctly reject the null 84 percent of the time, (incorrectly failing to reject 16 percent of the time).

- Test 2: Assume  $n = 25$  so our critical region could be defined as:

$$C = \{(x_1, \dots, x_n) : X_1 + X_2 + \dots + X_n > 25 * 78\} = \{(x_1, \dots, x_n) : \bar{X} > 78\}.$$

Thus,

$$\begin{aligned}
 K_2(\theta) &= Pr((x_1, \dots, x_n) \in C) \\
 &= Pr(\bar{X} \geq 78) \\
 &= Pr\left(\frac{\bar{X} - \theta}{2} \geq \frac{78 - \theta}{2}\right) \\
 &= Pr\left(Z \geq \frac{78 - \theta}{2}\right) \\
 &= 1 - Pr\left(Z \leq \frac{78 - \theta}{2}\right) \\
 &= 1 - N\left(\frac{78 - \theta}{2}\right)
 \end{aligned}$$

Note that:

$$K_2(73) = 0.006, \quad K_2(75) = 0.067, \quad K_2(79) = 0.681.$$

So if the true mean is 75, we now only (incorrectly) reject the null 6.7 percent of the time. This is better. However, if the mean is 79, we fail to reject the null (when it is false) over 30 percent of the time. Clearly there is a trade off between power - the

ability to reject a false null - and the ability not to reject when the null is true. Type I and II errors. G-23.2.

- Test 3: What if  $n$  was not given? Define:

$$C = \{(x_1, \dots, x_n) : \bar{X} > c\}.$$

And our power function:

$$\begin{aligned} K_3(\theta) &= Pr(\bar{X} \geq c) \\ &= Pr\left(\frac{\bar{X} - \theta}{10/\sqrt{n}} \geq \frac{c - \theta}{10/\sqrt{n}}\right) \\ &= Pr\left(Z \geq \frac{c - \theta}{10/\sqrt{n}}\right) \\ &= 1 - Pr\left(Z \leq \frac{c - \theta}{10/\sqrt{n}}\right) \\ &= 1 - N\left(\frac{c - \theta}{10/\sqrt{n}}\right) \end{aligned}$$

Suppose we would like to choose  $C$  and  $n$  such that  $K_3(75) = 0.159$  and  $K_3(77) = 0.841$ . This means:

$$\begin{aligned} K_3(75) = 0.159 &= 1 - N\left(\frac{c - \theta}{10/\sqrt{n}}\right) \Rightarrow \frac{c - \theta}{10/\sqrt{n}} = 1. \\ K_3(77) = 0.841 &= 1 - N\left(\frac{c - \theta}{10/\sqrt{n}}\right) \Rightarrow \frac{c - \theta}{10/\sqrt{n}} = -1. \end{aligned}$$

Two equations, two unknowns imply:  $n = 100$  and  $c = 76$ . We'll probably return to a situation like this later. It requires the ability to choose the size of our sample.

- **Definition:** Statistical Hypothesis: An assertion about the distribution of one or more random variables. We could have simple hypotheses like  $H_0 : \theta = 5$  or composite hypotheses like  $H_0 : \theta \leq 5$ .
- **Definition:** The power function is defined as:

$$K_{\Upsilon}(\theta) = Pr((x_1, \dots, x_n) \in C_{\Upsilon}).$$

The probability of a rejection for a given value of the true parameter,  $\theta$ .

- **Definition:** Significance Level. Define the parameter space as  $\Theta$  with:

$$\Theta = \Theta_0 \cup \Theta_0^c,$$

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_0^c.$$

Then,

$$\alpha = \text{Sup}_{\theta \in \Theta_0} K(\theta),$$

is the significance level of the test. It is the maximum probability of rejecting when the null hypothesis is true. See graph G-23.3.

## 24 Lecture 24: November 30, 2004

### 24.1 More on Hypothesis Testing

- See G-24.1. Consider our usual test:

$$H_0 : \theta \in \Theta_0, \text{ vs } H_1 : \theta \in \Theta - \Theta_0.$$

Denote  $K(\theta)$  the power function, or the probability of rejecting  $H_0$  given  $\theta$ . Then, given the points  $A$  and  $B$  in the graph,

$$K(A) = Pr(\text{Type I Error}) \equiv \text{Reject } H_0 \text{ when it is true.}$$

$$1 - K(B) = Pr(\text{Type II Error}) \equiv \text{Fail to Reject } H_0 \text{ when it is false.}$$

Note that the maximum probability of a type I error is called the significance level of the test.

### 24.2 Most Powerful Tests

- Consider a parameter space consisting of only two possible values:

$$\Theta = \{\theta', \theta''\}.$$

Hypotheses:

$$H_0 : \theta = \theta', \text{ vs } H_1 : \theta = \theta''.$$

Suppose we have two possible tests,  $\Upsilon$  and  $\Upsilon^*$ . Denote:

$$K_{\Upsilon}(\theta'),$$

the significance level of the test ( $\Pr(\text{type I error})$ ). Also, denote:

$$1 - K_{\Upsilon}(\theta''),$$

the probability of a type II error. Finally,

$$K_{\Upsilon^*}(\theta'),$$

the significance level of the second test.

- **Definition:** A most powerful test. Let  $K_{\Upsilon^*}(\theta') = \alpha$ . Then if,

$$K_{\Upsilon^*}(\theta'') \geq K_{\Upsilon}(\theta'') \forall \text{ tests, } \Upsilon, \text{ with, } K_{\Upsilon}(\theta') = \alpha,$$

Then  $\Upsilon^*$  is the most powerful test. Hence we fix a significance level and then find the test which minimizes the probability of a type II error.

- **Definition:** A Simple Likelihood Ratio Test. Assume we have a random sample,  $X_1, \dots, X_n$ , with density,  $f(x; \theta)$ , where  $\theta \in \{\theta', \theta''\}$ . Hypotheses:

$$H_0 : \theta = \theta', \text{ vs } H_1 : \theta = \theta''.$$

A test  $\Upsilon$  defined by:

$$\text{Reject } H_0 \text{ if: } \lambda \leq k,$$

$$\text{Do Not Reject } H_0 \text{ if: } \lambda > k.$$

Where,

$$\lambda = \lambda(X_1, \dots, X_n) = \frac{\prod_{i=1}^n f(x_i, \theta')}{\prod_{i=1}^n f(x_i, \theta'')},$$

is called a Simple Likelihood Ratio Test (simple LRT). See G-24.2 for a graphical interpretation.

- **Theorem:** Neijman - Pearson. Denote hypotheses:

$$H_0 : \theta = \theta', \text{ vs } H_1 : \theta = \theta''.$$

Likelihood function:

$$L(\theta, x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta).$$

Denote the critical region for the test:

$$C = \left\{ (x_1, \dots, x_n) : \frac{L(\theta', x_1, \dots, x_n)}{L(\theta'', x_1, \dots, x_n)} \leq k \right\}.$$

With,

$$Pr((x_1, \dots, x_n) \in C; H_0) = \alpha.$$

Then the test defined by the critical region,  $C$ , is the most powerful test. Any other test (any other critical region) with significance level,  $\alpha$ , would be less powerful.

- **Example.** Suppose  $X_1, \dots, X_n \sim N(\theta, 1)$ , with  $\theta \in \{\theta', \theta''\} = \{0, 1\}$ . Hypotheses:

$$H_0 : \theta = 0, \text{ vs } H_1 : \theta = 1.$$

Density:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5(x-\theta)^2}.$$

Likelihood Ratio:

$$\begin{aligned}
 \frac{L(\theta', x_1, \dots, x_n)}{L(\theta'', x_1, \dots, x_n)} &= \frac{\prod_{i=1}^n f(x_i, \theta')}{\prod_{i=1}^n f(x_i, \theta'')} \\
 &= \frac{\prod_{i=1}^n f(x_i, 0)}{\prod_{i=1}^n f(x_i, 1)} \\
 &= \frac{e^{-0.5 \sum (x_i - 0)^2}}{e^{-0.5 \sum (x_i - 1)^2}} \\
 &= e^{-0.5 \sum (x_i)^2 + 0.5 \sum (x_i - 1)^2} \\
 &= e^{-0.5 \sum (x_i)^2 + 0.5 \sum x_i^2 - \sum x_i + 0.5 * n} \\
 &= e^{-\sum x_i + 0.5 * n}
 \end{aligned}$$

So we could define our critical region to be:

$$C = \left\{ (x_1, \dots, x_n) : e^{-\sum x_i + 0.5 * n} \leq k \right\}.$$

But we could also manipulate the expression on the right to be:

$$\begin{aligned}
 e^{-\sum x_i + 0.5 * n} &\leq k \\
 -\sum x_i + 0.5 * n &\leq \ln(k) \\
 \sum x_i &\geq 0.5 * n - \ln(k) \equiv c \\
 \sum x_i &\geq c \\
 (1/n) \sum x_i &\geq (1/n)c \equiv c_1 \\
 \bar{x} &\geq c_1
 \end{aligned}$$

So now our critical region becomes:

$$C = \left\{ (x_1, \dots, x_n) : \bar{x} \geq c_1 \right\}.$$

- How do we choose  $c_1$ ? Lets choose a significance level,  $\alpha = 0.05$ . Now we would like to have  $c_1$  such that:

$$Pr((x_1, \dots, x_n) \in C; H_0) = K(\theta') = K(0) = \alpha = 0.05.$$

Note that under the null,  $X_i \sim N(0, 1)$  so  $\bar{X} \sim N(0, 1/n)$ . Assume  $n = 25$ . Thus,

$$\begin{aligned}
 Pr((x_1, \dots, x_n) \in C; H_0) &= Pr(\bar{x} \geq c_1; H_0) \\
 &= Pr(\underbrace{\sqrt{n}\bar{x}}_{\sim N(0,1)} \geq \sqrt{n}c_1; H_0) \\
 &= 1 - N(\sqrt{n}c_1) = 0.05 \\
 \Rightarrow 5 * c_1 &= 1.645 \\
 \Rightarrow c_1 &= 0.329
 \end{aligned}$$

- So our critical region is finally,

$$C = \left\{ (x_1, \dots, x_n) : \bar{x} \geq 0.329 \right\}.$$

So the probability that a set of realizations of  $X$  will belong to  $C$  is 0.05. If  $X \in C$ , then reject the null.

- Finally, we can determine the the power of the test we just formulated now that we have  $c_1$  defined. Thus,

$$\begin{aligned}
 K(\theta'') = K(1) &= Pr(\bar{x} \geq c_1; H_1) \\
 &= Pr(\bar{x} - 1 \geq c_1 - 1; H_1) \\
 &= Pr(\underbrace{\sqrt{n}(\bar{x} - 1)}_{\sim N(0,1)} \geq \sqrt{n}(c_1 - 1); H_1) \\
 &= 1 - N(\sqrt{n}(c_1 - 1)) \\
 &= 1 - N(5(0.329 - 1)) \\
 &= 1 - N(-3.3) \approx 1
 \end{aligned}$$

So the power of the test is:  $1 - K(\theta'') \approx 0$ , damn good.

## 25 Lecture 25: December 2, 2004

### 25.1 Uniformly Most Power Tests

- **Definition:** Let the parameter space be such that:

$$H_0 : \theta = \theta' \text{ vs } H_1 : \theta \in \Theta - \Theta_0.$$

So maybe  $\Theta = [0, \infty)$ ,  $\theta' = 0$ ,  $\Theta - \{0\} = (0, \infty)$ . So we have a simple null hypothesis and a composite alternative. Consider a test,  $\Upsilon^*$ , with critical region,  $C^*$ , such that:

$$K_{\Upsilon^*}(\theta') = Pr((x_1, \dots, x_n) \in C^*; H_0) = \text{Significance Level} = \alpha.$$

Then if:

$$K_{\Upsilon^*}(\theta) \geq K_{\Upsilon}(\theta), \forall \Upsilon \ni K_{\Upsilon}(\theta') = \alpha, \forall \theta \in \Theta - \theta',$$

then  $\Upsilon^*$  is the Uniformly Most Powerful Test (UMPT).

- See G-25.1. The idea is that we fix the significance level of the test, but instead of just choosing the test with the smallest probability of a type II error for a given  $\theta$ , we look over all tests with the same significance level and ALSO over all values of  $\theta$  in the alternative hypothesis set. Clearly, a UMPT may not exist.
- Suppose  $H_0 : \theta = \theta'$  and  $H_1 : \theta = \theta''$  where  $\theta'' \in \Theta - \{\theta'\}$ . Here we have a simple null versus a simple alternative so a most powerful test will exist by the Neijman - Pearson (NP) theorem. If the critical region of the MPT does not depend on  $\theta''$ , then this MPT is also a UNIFORMLY MPT.
- Example. Suppose  $X_i \sim N(0, \theta)$ . Parameter space:  $\Theta = \{\theta : \theta \geq \theta'\}$ . Hypotheses:

$$H_0 : \theta = \theta', \text{ vs } H_1 : \theta > \theta'.$$

This is a composite hypothesis, but lets try to reformulate the problem into a simple alternative and then see if we could apply it to this composite test. Consider:

$$H_0 : \theta = \theta', \text{ vs } H_1 : \theta = \theta'', \text{ where } \theta'' > \theta.$$

Here we have two simple hypotheses, so via the NP theorem, the most powerful test is a likelihood ratio test. Consider the likelihood functions:

$$L(\theta'; x_1, \dots, x_n) = \prod f(x_i, \theta') = \frac{1}{(2\pi\theta')^{n/2}} e^{-\frac{1}{2\theta'} \sum x_i^2}.$$

$$L(\theta''; x_1, \dots, x_n) = \prod f(x_i, \theta'') = \frac{1}{(2\pi\theta'')^{n/2}} e^{-\frac{1}{2\theta''} \sum x_i^2}.$$

So the critical region is defined by:

$$\begin{aligned} \frac{L(\theta', x_1, \dots, x_n)}{L(\theta'', x_1, \dots, x_n)} &\leq k \\ \frac{1}{(2\pi\theta')^{-n/2} e^{-\frac{1}{2\theta'} \sum x_i^2}} &\leq k \\ \frac{1}{(2\pi\theta'')^{-n/2} e^{-\frac{1}{2\theta''} \sum x_i^2}} &\leq k \\ \left(\frac{\theta''}{\theta'}\right)^{n/2} e^{0.5 \sum x_i^2 (1/\theta'' - 1/\theta')} &\leq k \\ \left(\frac{\theta''}{\theta'}\right)^{n/2} e^{-\frac{\theta'' - \theta'}{2\theta'\theta''} \sum x_i^2} &\leq k \\ \frac{n}{2} \log\left(\frac{\theta''}{\theta'}\right) - \frac{\theta'' - \theta'}{2\theta'\theta''} \sum x_i^2 &\leq \log(k) \\ \frac{n}{2} \log\left(\frac{\theta''}{\theta'}\right) - \log(k) &\leq \frac{\theta'' - \theta'}{2\theta'\theta''} \sum x_i^2 \\ \text{Because } \theta'' - \theta' > 0\dots & \\ \frac{2\theta'\theta''}{\theta'' - \theta'} \left(\frac{n}{2} \log\left(\frac{\theta''}{\theta'}\right) - \log(k)\right) &\leq \sum x_i^2 \\ \sum x_i^2 &\geq c \end{aligned}$$

- So we will choose  $c$  to attain  $\alpha = 0.05$ . The critical region is defined by:

$$C = \{(x_1, \dots, x_n) : \sum x_i^2 \geq c\}.$$

So set  $c$  such that,

$$\alpha = Pr((x_1, \dots, x_n) \in C; H_0) = Pr(\sum x_i^2 \geq c; H_0) = Pr\left(\frac{\sum x_i^2}{\theta'} \geq \frac{c}{\theta'}; H_0\right).$$

Note that  $\frac{X_i}{\sqrt{\theta'}} \sim N(0, 1)$ . Thus  $\sum_i \frac{X_i^2}{\theta'} \sim \chi^2(n)$ . So,

$$\alpha = Pr\left(\frac{\sum x_i^2}{\theta'} \geq \frac{c}{\theta'}; H_0\right) = Pr(\chi^2(n) \geq \frac{c}{\theta'}).$$

Suppose  $n = 15$ ,  $\alpha = 0.05$  and  $\theta' = 3$ . Then:

$$Pr(\chi^2(15) \geq \frac{c}{3}) = 0.05 \implies c = 75.$$

So our critical region becomes:

$$C = \{(x_1, \dots, x_{15}) : \sum x_i^2 \geq 75\}.$$

Note that this is true for  $\theta'' = 4$ ,  $\theta'' = 54.2$ , etc. It's true for ANY value of  $\theta'' > 3$ . So we have found a critical region for all possible  $\theta \in \Theta - \theta'$ . Hence our test is uniformly most powerful. Or:

$$K_{\Upsilon^*}(\theta') = 0.05, \text{ and } K_{\Upsilon^*}(\theta'') \geq K_{\Upsilon}(\theta'') \forall \Upsilon, \forall \theta''.$$

## 25.2 General Likelihood Ratio Test

- **Definition:** Suppose  $X_1, \dots, X_n$  is a random sample with density  $f(x, \theta)$  where  $\theta = (\theta_1, \dots, \theta_m) \in \Theta$ . Hypotheses:

$$H_0 : \theta \in \Theta_0, \text{ vs } H_1 : \theta \in \Theta - \Theta_0.$$

So maybe  $H_0 : \theta_1 + \theta_3 \leq 5$ . Following the same lines as in the simple likelihood ratio test,

$$\lambda(x_1, \dots, x_n) = \frac{\text{Sup}_{\theta \in \Theta_0} L(\theta; x_1, \dots, x_n)}{\text{Sup}_{\theta \in \Theta} L(\theta; x_1, \dots, x_n)}.$$

See G-25.2. If the top is much smaller than the bottom,  $\lambda$  will be close to 0 and clearly the restriction is having a large impact on the likelihood function. If it is near 1, the null hypothesis may be ok. Thus if:

$$\lambda \leq \lambda_0,$$

reject  $H_0$ . We could also write the statistic as:

$$\tilde{\lambda} = \frac{L(\hat{\theta}_{ML}^{\text{Restricted}}, x_1, \dots, x_n)}{L(\hat{\theta}_{ML}^{\text{Unrestricted}}, x_1, \dots, x_n)}.$$

Which in reality is what we will do to run this test. What is  $\lambda_0$ ? Choose  $\lambda_0$  such that:

$$\text{Sup}_{\theta \in \Theta_0} \underbrace{\text{Pr}_{\theta}(\lambda(x_1, \dots, x_n) \leq \lambda_0)}_{K_{LR}(\theta)} = \alpha \equiv \text{Desired Significance Level}.$$

So if we want a significance level of 5%, when we just set  $\lambda_0$  such that the maximum probability of a type I error over all  $\theta$  under the null is equal to 0.05.

- **Example.** Suppose  $X_1, \dots, X_n$  is a random sample such that  $X_i \sim N(\theta_1, \theta_2)$ . Suppose our parameter space is:

$$\Theta = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}.$$

Hypotheses:

$$H_0 : \theta_1 = 0, \theta_2 > 0, \quad vs \quad H_1 : \theta_1 \neq 0, \theta_2 > 0.$$

No restriction on  $\theta_2$ . Note our null is now composite. Then our likelihood function is:

$$L(\theta_1, \theta_2; x_1, \dots, x_n) = (2\pi\theta_2)^{-n/2} e^{-0.5\theta_2 \sum (x_i - \theta_1)^2}.$$

Now we have five steps:

- (1) Maximize  $L(\cdot)$  over  $\theta \in \Theta$ .

$$\text{Max}_{\theta \in \Theta} L(\theta_1, \theta_2; x_1, \dots, x_n).$$

Setting both partials equal to 0 yields:

$$\hat{\theta}_1 = \frac{1}{n} \sum x_i,$$

$$\hat{\theta}_2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

- (2) Maximize  $L(\cdot)$  over  $\theta \in \Theta_0$ .

$$\text{Max}_{\theta \in \Theta_0} L(0, \theta_2; x_1, \dots, x_n) = (2\pi\theta_2)^{-n/2} e^{-0.5\theta_2 \sum x_i^2}.$$

Setting the partial equal to 0 yields:

$$\tilde{\theta}_2 = \frac{1}{n} \sum x_i^2.$$

- (3) Plug the maximized values back into the Unrestricted likelihood function.

$$\begin{aligned} L(\hat{\theta}_1, \hat{\theta}_2; x_1, \dots, x_n) &= \left( \frac{1}{2\pi \frac{1}{n} \sum (x_i - \bar{x})^2} \right)^{n/2} e^{-\frac{1}{2} \frac{\sum (x_i - \bar{x})^2}{\frac{1}{n} \sum (x_i - \bar{x})^2}} \\ &= \left( \frac{ne^{-1}}{2\pi \sum (x_i - \bar{x})^2} \right)^{n/2}. \end{aligned}$$

- (4) Plug the maximized values back into the Restricted likelihood function.

$$\begin{aligned} L(0, \tilde{\theta}_2; x_1, \dots, x_n) &= \left( \frac{1}{2\pi \frac{1}{n} \sum x_i^2} \right)^{n/2} e^{-\frac{1}{2} \frac{\sum x_i^2}{\frac{1}{n} \sum x_i^2}} \\ &= \left( \frac{ne^{-1}}{2\pi \sum x_i^2} \right)^{n/2}. \end{aligned}$$

– (5) Compute the ratio of restricted over unrestricted:

$$\lambda = \frac{\sum (x_i - \bar{x})^2}{\sum x_i^2} = \frac{1}{\left(1 + \underbrace{\frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2}}_A\right)^{n/2}} \leq \lambda_0.$$

$$\frac{1}{(1 + A)^{n/2}} \leq \lambda_0.$$

$$1 + A \geq \lambda_0^{-n/2}.$$

$$A \geq \lambda_0^{-n/2} - 1 = c.$$

$$\frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2} \geq c.$$

$$\frac{\sqrt{n}|\bar{x}|}{\sqrt{\sum (x_i - \bar{x})^2}} \geq \sqrt{c}.$$

$$\underbrace{\frac{\sqrt{n}|\bar{x}|}{\sqrt{\sum (x_i - \bar{x})^2 / (n - 1)}}}_{|t|} \geq \sqrt{c / (n - 1)} = c_0.$$

- So in the final step, we transformed the probability into something that looks like a  $t$ . We can then choose a  $c_0$  so that this probability is equal to our desired significance level. Note the test is now two sided so we would be interested in the 97.5% fractile if we wanted  $\alpha = 0.05$ .

## 26 Lecture 26: December 9, 2004

### 26.1 Asymptotic Theory

- Consider the sample mean of a random variable:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Recall our definition of a limit:

$$\lim_{n \rightarrow \infty} a_n = a \iff \text{if } \forall \epsilon \exists N_\epsilon \ni |a_n - a| < \epsilon \forall n \geq N_\epsilon.$$

- So what about  $\lim_{n \rightarrow \infty} \hat{\theta}_n$  ?

#### Weak Convergence in Probability

- **Definition:** Let  $Z_n = \hat{\theta}_n$ , our estimator. Then if:

$$\lim_{n \rightarrow \infty} \Pr(|Z_n - Z| \leq \epsilon) = 1 \quad \forall \epsilon > 0,$$

we say that  $Z_n$  converges in probability to  $Z$ . Or,

$$p\lim_{n \rightarrow \infty} Z_n = Z, \text{ or } Z_n \xrightarrow{p} Z, \text{ or } Z_n \rightarrow Z \text{ i.p. as } n \rightarrow \infty.$$

#### Strong Convergence in Probability (Almost Surely)

- Consider a probability space  $(\Omega, \mathfrak{A}, P)$  with  $Z : \Omega \mapsto \mathfrak{R}$  and  $Z_n : \Omega \mapsto \mathfrak{R}$ . Then if

$$Z_n(\omega) \rightarrow Z(\omega) \quad \forall \omega \in \Omega - N, \text{ where } P(N) = 0,$$

we say that  $Z_n$  almost surely converges to  $Z$  in probability. This type of convergence is stronger than weak convergence since it is true for all realizations,  $\omega$ . The set  $N$  is difficult to define but it might be like the irrationals on the real line (more on this next semester). We write:

$$Z_n \xrightarrow{a.s.} Z.$$

#### Convergence in the $r^{\text{th}}$ Mean

- $Z_n$  converges in the  $r^{\text{th}}$  mean to  $Z$  if:

$$E[|Z_n - Z|^r] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

If  $E[|Z_n - Z|^2] \rightarrow 0$ , then we say that  $Z_n$  converges in the quadratic mean to  $Z$ .

- Two implications:

Strong (Almost Sure) Convergence  $\implies$  Weak Convergence.

$r^{\text{th}}$  Mean Convergence  $\implies$  Weak Convergence.

We can't say anything comparing strong and  $r^{\text{th}}$  mean convergence.

- Proof that  $r^{\text{th}}$  mean convergence implies weak convergence. Consider:

$$\begin{aligned} Pr(|Z_n - Z| \geq \epsilon) &= Pr(|Z_n - Z|^r \geq \epsilon^r) \\ &\leq \frac{E[|Z_n - Z|^r]}{\epsilon^r} \longrightarrow 0 \end{aligned}$$

So  $r^{\text{th}}$  mean convergence implies weak convergence.

- Corollary. If  $E[Z_n] = Z$  and  $Var(Z_n) \rightarrow 0$ , then  $Z_n \rightarrow^p Z$ . We showed this before for consistent estimators - must be unbiased and have zero variance in the limit.
- Example. Suppose  $X_i \sim \text{iid}(\mu, \sigma^2)$ . Let:

$$Z_n = \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Thus,

$$Z = \mu.$$

- Weak Convergence? Yes:

$$E[Z_n] = \mu, \lim_{n \rightarrow \infty} Var(Z_n) = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0 \Rightarrow Z_n \rightarrow^p Z.$$

- Quadratic Mean Convergence? Yes:

$$E[(Z_n - Z)^2] = E[(Z_n - \mu)^2] = \frac{\sigma^2}{n} \Rightarrow \lim_{n \rightarrow \infty} E[(Z_n - Z)^2] = 0 \Rightarrow Z_n \rightarrow^{q.m.} Z.$$

- Strong Convergence? Yes. Not shown.

- Now suppose  $X_i \sim \text{iid}$  with  $E[X] = \mu$ , variance unspecified. Then, it can be shown,

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow^{a.s.} \mu.$$

So we don't need the existence of a variance to get strong convergence. We know this implies weak convergence but it is also clear that it does NOT necessarily imply convergence in the quadratic mean since the variance may not exist!

- For  $n$ -dimensional random variables, you simply need to look at each component and overall convergence (of any type) is implied by the convergence of each component.

## 26.2 Convergence in Distribution

- Suppose we have distribution functions  $F_1, F_2, \dots$ . If we can show that  $F_n(z) \rightarrow F(z)$  then we can say that  $F_n$  converges in distribution to  $F$ .
- Consider  $X_i \sim \text{iid } N(0, \sigma^2)$ . Then,

$$\hat{\theta}_n = \frac{1}{n} \sum X_i \sim N(0, \sigma^2/n).$$

Let  $Z_n = \hat{\theta}_n$  with corresponding distribution function  $F_n(z)$ . Define another distribution:

$$F(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases}$$

See G-26.1. We plot the CDF,  $F_n(z)$  for various levels of  $n$ . Note for:

$$z > 0 \Rightarrow F_n(z) \rightarrow 1,$$

$$z < 0 \Rightarrow F_n(z) \rightarrow 0,$$

$$z = 0 \Rightarrow F_n(z) \rightarrow 1/2.$$

So,

$$F_n(z) \rightarrow F(z) \quad \underbrace{\forall z \text{ except } z = 0}_{\forall z \text{ where } F(z) \text{ is continuous}}.$$

And this is our definition of convergence in distribution. Note:

$$\lim_{n \rightarrow \infty} F_n(z) = F(z) \iff \lim_{n \rightarrow \infty} \Pr(Z_n \leq z) = \Pr(Z \leq z).$$

Or the probability law of  $Z_n$  converges to the probability law of  $Z$ .

- Finally consider  $X_i \sim \text{iid } (\theta, \sigma^2)$ . Then,

$$\hat{\theta}_n = \frac{1}{n} \sum X_i \sim (\theta, \sigma^2/n).$$

Let  $Z_n = \sqrt{n}(\hat{\theta}_n - \theta)$ . Then,

$$E[Z_n] = 0.$$

$$\text{Var}[Z_n] = n * \text{Var}(\hat{\theta}_n) = \sigma^2.$$

So,

$$\hat{\theta}_n = \theta + \frac{1}{\sqrt{n}} Z_n \sim N\left(\theta, \frac{\sigma^2}{n}\right).$$

This result comes from the central limit theorem.

# Final Review

## 26.3 Key Lecture Notes - Part I

- $\mathfrak{A}$  is a  $\sigma$ -algebra if: (1)  $\emptyset \in \mathfrak{A}$ ,  $\Omega \in \mathfrak{A}$ , (2) if  $A \in \mathfrak{A}$  then  $A^c \in \mathfrak{A}$ , and (3) Infinite unions are in.
- Properties of a (probability) measure : (1)  $P(\emptyset) = 0$ ,  $P(\Omega) = 1$ , (2)  $P(A) \geq 0$ , and (3)  $P(\cup_i A_i) = \sum_i P(A_i)$  if  $A$ 's disjoint.
- $X : \Omega \mapsto \Omega^+$ ,  $X$  is  $(\mathfrak{A}, \mathfrak{A}^+)$ -measurable if:

$$X^{-1}(A^+) \in \mathfrak{A} \forall A^+ \in \mathfrak{A}^+.$$

- Properties of distribution functions: (1)  $F(x)$  nondecreasing, (2)  $F(\infty) = 1$ ,  $F(-\infty) = 0$ , and (3)  $F(x)$  is right-continuous.
- Important theorem: Let  $F(x)$  be a distribution function. Then  $\exists$  a unique probability measure:

$$P_*((a, b]) = F(b) - F(a) \geq 0, \text{ for } a < b.$$

There is a 1:1 relationship between the probability law for a random variable,  $P_x$ , and the distribution function of  $X$ ,  $F(x)$ .

- CDF and PDF:

$$F(x) = Pr(X \leq x) = \sum_{x^i \in (-\infty, x)} f(x^i) \text{ OR } \int_{-\infty}^x f(x) dx.$$

$$f(x) = Pr(X = x) = F(x) - F(x-).$$

- Properties of a PDF: (1)  $0 \leq f(x) \leq 1$ , and (2)  $\sum_x f(x) = 1$  OR  $\int_{-\infty}^{\infty} f(x) dx = 1$ .
- CDF to PDF:  $f(x) = \frac{\partial F(x)}{\partial x}$ . Thus the PDF is not unique for a continuous RV.

- Expectation:

$$E[u(x)] = \sum_{x^i} u(x^i) f(x^i) \text{ OR } \int_{-\infty}^{\infty} u(x) f(x) dx.$$

- In general:  $E[u(x, y)] \neq u(E[x], E[y])$ .
- $E[(X - a)^r]$  is the  $r^{th}$  central moment around  $a$ .

$$\text{Variance} : E[(X - E[X])]^2 = \sigma_x^2 = E[X^2] - \mu_x^2.$$

$$\text{Skewness} : E[(X - E[X])]^3 / \sigma_x^3.$$

$$\text{Excess Kurtosis} : E[(X - E[X])]^4 / \sigma_x^4 - 3.$$

- MGF:  $M(t) = E[e^{tX}]$ .  $E[X^k] = \left. \frac{\partial^k M(t)}{\partial t^k} \right|_{t=0}$

- Chebyshev's Inequality:

$$Pr[u(x) \geq \epsilon] \leq \frac{E[u(x)]}{\epsilon} \text{ OR } Pr[|X - \mu_x| \geq k\sigma_x] \leq \frac{1}{k^2}.$$

- Conditional Probability:  $P(A|B) = P(A \cap B)/P(B)$ .

- Total Probability:

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

- Bayes:

$$P(B_k|A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A|B_k) * P(B_k)}{\sum_{i=1}^n P(A|B_i) * P(B_i)}.$$

- Independence of Events:  $P(A \cap B) = P(A) * P(B)$ ,  $P(A|B) = P(A)$ ,  $P(B|A) = P(B)$ .

- Given the joint distribution  $f(x, y)$ ,

$$f_x(x) = \sum_y f(x, y) \text{ OR } \int_{-\infty}^{\infty} f(x, y) dy.$$

- MGF of a vector RV:  $M(t_1, \dots, t_n) = E[e^{t_1x_1 + t_2x_2 + \dots + t_nx_n}]$ .

- Conditional distribution:

$$f(x|y) = \frac{f(x, y)}{f(y)}.$$

- Iterative Expectations:

$$E[E[X|Y]] = E[X].$$

- Independence of RVs:  $f(x|y) = f(x)$ ,  $f(x, y) = f(x) * f(y)$ ,  $E[XY] = E[X] * E[Y]$ ,  $M_{xy}(t_1, t_2) = M_x(t_1) * M_y(t_2)$ ,  $E[X|Y = y] = E[X]$ .

- Functions of independent random variables are independent.

- Covariance:  $\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$ . Linear Relationship.

- Correlation Coefficient:  $\rho_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$ .

- Sample Mean and Variance:

$$\bar{X} = \frac{1}{n} \sum_i X_i.$$

$$S^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2.$$

$$S_1^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2.$$

- Variance of the sample mean of a RANDOM sample:  $\frac{\sigma^2}{n}$ .

## 26.4 Key Lecture Notes - Part II

- Change of variables for continuous RVs:  $X$  is a RV.  $Y = u(X)$ . Then:

$$g(y) = f(w(y)) * |\partial w(y)/\partial y| \text{ for } y \in \mathcal{Y}.$$

- $Z \sim N(0, 1)$ ,  $Y_2 \sim \chi^2(r_2)$  with  $Z$  and  $Y_2$  independent, then  $X = Z/\sqrt{Y_2/r_2} \sim t(r_2)$ .
- $Y_1 \sim \chi^2(r_1)$ , independence, then  $X = (Y_1/r_1)/(Y_2/r_2) \sim F(r_1, r_2)$ .
- Note  $Z^2 = Y_1/1$  so

$$\frac{Z^2}{Y_2/r_2} \sim F(1, r_2) \text{ and } \frac{Z}{\sqrt{Y_2/r_2}} \sim t(r_2)$$

- MGF Technique.  $Y = u(X)$ . Then:

$$M_y = E[e^{t_1 y_1 + t_2 y_2 + \dots + t_s y_s}] = \int e^{t_1 u_1(x_1, \dots, x_n) + \dots + t_n u_n(x_1, \dots, x_n)} f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

- If  $X_i \sim \chi^2(r_i)$  independent, then  $\sum_{i=1}^n X_i \sim \chi^2(r)$  where  $r = \sum_{i=1}^n r_i$ .
- If you start with a normal, standardize it, you get a  $N(0, 1)$ . Square it, you get a  $\chi^2(1)$ . Add a bunch of these, you get a  $\chi^2(n)$ .
- $X \sim N(\mu, \Sigma)$ .  $X$  has MGF  $M(t) e^{\mu' t + 0.5 t' \Sigma t}$ .
- $X \sim N(\mu, \Sigma)$  then  $X_i \sim N(\mu_i, \sigma_{ii})$ .
- For normally distributed random variables Zero Covariance Implies Independent (ONLY for Normals!)
- If  $Z \sim N(0, \sigma^2 I)$ , then the quadratic form  $Q = (Z' A Z)/\sigma^2$  and the linear form,  $X = B Z/\sigma$  are independent if  $BA = 0$ .
- Two quadratic forms  $Q = Z' A Z$  and  $R = Z' B Z$  are independent if  $BA = 0$ .
- Sample variance (biased and unbiased):

$$S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2, \quad S_1^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2.$$

- Since  $\bar{X}$  is a linear form and  $nS^2$  is a quadratic form then the sample mean and sample variance are independent.
- If  $X_i \sim (\mu, \sigma^2)$ , then  $\bar{X} \approx N(\mu, \sigma^2/n)$  by the CLT.
- If  $X_i \sim iid N(\mu, \sigma^2)$ , then  $\sqrt{n-1} \frac{\bar{X} - \mu}{S} \sim t(n-1)$ .

- Maximum Likelihood Estimator:  $\hat{\mu} = arg \max_{\mu} f(X, \mu)$ . The  $\hat{\mu}$  as a function of realizations is an estimate. As a function of random variables, it is an estimator (also a random variable). Note:

$$L[u(x_1, \dots, x_n); x_1, \dots, x_n] \geq L[\theta; x_1, \dots, x_n].$$

So the ML estimator is not necessarily unique.

- The ML estimator of the variance of a normal is biased.

- Unbiased:

$$E_{\theta_0}[\hat{\theta}] = \theta_0 \forall \theta_0 \in \Theta.$$

- Consistent:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0 \forall \epsilon > 0.$$

- MSE:

$$E[(\hat{\theta}_n - \theta)^2] = Var(\hat{\theta}_n) + [Bias(\hat{\theta}_n)]^2.$$

Sufficient conditions for consistence is unbiasedness in the limit and 0 variance in the limit. Each alone is not enough.

- Method of Moments: Set population moments equal to sample moments and solve for your parameters.
- You need at least as many (independent) moments as you have parameters to estimate. The ML and MM estimators correpond exactly to the OLS estimators.
- Unbiased Minimum Variance Estimator: Choose the lowest variance estimator among those that are unbiased. So we are minimizing the MSE but since the bias is zero, we just have the variance term. Ideally, we would choose a technique for each problem because of the particulars. Maybe weighting positive deviations more, etc.
- Confidence Intervals: A  $\gamma$  percent CI for the sample mean when the variance is known is  $\bar{X} \pm z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}$ .
- Confidence Intervals: A  $\gamma$  percent CI for the sample mean when the variance is unknown is  $\bar{X} \pm t_{(\gamma+1)/2}(n-1) \frac{S}{\sqrt{n-1}}$ .
- Critical Region,  $C$ . If the realization fall into  $C$ , then reject. This region completely defines the test.

- $K(\theta)$  is the power function: The probability of rejecting the null for a given value of the true parameter  $\theta$ .

$$K_{\Upsilon}(\theta) = Pr((x_1, \dots, x_n) \in C_{\Upsilon}).$$

- Significance Level:  $\alpha = \text{Sup}_{\theta \in \Theta_0} K(\theta)$ , ie the maximum probability of rejection when the Null is true.
- Type I Error: Reject when  $H_0$  is true. Type II Error: Fail to Reject when  $H_0$  is false.
- Most Powerful Test. Fix  $\alpha$  and then choose the test with the highest power among those with a significance level of  $\alpha$ .

$$K_{\Upsilon^*}(\theta'') \geq K_{\Upsilon}(\theta'') \quad \forall \text{ tests, } \Upsilon, \text{ with, } K_{\Upsilon}(\theta') = \alpha.$$

- Simple Likelihood Ratio Test. Simple null  $H_0 : \theta = \theta'$  versus  $H_1 : \theta = \theta''$ .

$$\lambda = \frac{\prod_{i=1}^n f(x_i; \theta')}{\prod_{i=1}^n f(x_i; \theta'')}.$$

Reject  $H_0$  if  $\lambda \leq k$ .

- Neijman Pearson - For a simple null and alternative, the simple likelihood ratio test is the most powerful test.
- Uniformly Most Power Test. Fix  $\alpha$  and then choose the test with the highest power among those with a significance level of  $\alpha$  but also search over all values of  $\theta \in \Theta - \Theta_0$ .

$$K_{\Upsilon^*}(\theta) \geq K_{\Upsilon}(\theta), \quad \forall \Upsilon \ni K_{\Upsilon}(\theta') = \alpha, \quad \forall \theta \in \Theta - \theta'.$$

- General Likelihood Ratio Test - Composite hypotheses.

$$\lambda = \frac{\text{Sup}_{\theta \in \Theta_0} L(\theta; x_1, \dots, x_n)}{\text{Sup}_{\theta \in \Theta} L(\theta; x_1, \dots, x_n)}.$$

Reject null if  $\lambda \leq k$ . Often write statistic as:

$$\tilde{\lambda} = \frac{L(\hat{\theta}_{ML}^{Restricted}; x_1, \dots, x_n)}{L(\hat{\theta}_{ML}^{Unrestricted}; x_1, \dots, x_n)}.$$

Steps:

- (1) Maximize  $L(\cdot)$  over  $\theta \in \Theta$ .
- (2) Maximize  $L(\cdot)$  over  $\theta \in \Theta_0$ .
- (3) Plug the maximized values back into the Unrestricted likelihood function.
- (4) Plug the maximized values back into the Restricted likelihood function.
- (5) Compute the ratio of restricted over unrestricted:

## 26.5 Discrete Distributions

### Bernoulli

- PDF:

$$f(x) = p^x(1-p)^{1-x} \text{ for } x = 0, 1.$$

- Mean:  $p$ .
- Variance:  $p(1-p)$ .
- Success or Failure.

### Binomial

- PDF:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \dots, n.$$

- Mean:  $np$ .
- Variance:  $np(1-p)$ .
- Number of successes in  $n$  trials. Sum of iid Bernoulli trials.

### Poisson

- PDF:

$$f(x) = \frac{e^{-m} m^x}{x!} \text{ for } x = 0, 1, \dots$$

- Mean:  $m$ .
- Variance:  $m$ .
- Binomial with large  $n$  and small  $p$ .

### Geometric

- PDF:

$$f(x) = p(1-p)^x \text{ for } x = 0, 1, \dots$$

- Mean:  $(1-p)/p$ .
- Variance:  $(1-p)/p^2$ .
- Number of Bernoulli trials before the first success.

## 26.6 Continuous Distributions

### Continuous Uniform

- PDF:

$$f(x) = \frac{1}{b-a}, \text{ for } a \leq x \leq b.$$

- Mean:  $(a+b)/2$ .
- Variance:  $(b-a)^2/12$ .

### Normal

- PDF:

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

- Mean:  $\mu$ .
- Variance:  $\sigma^2$ .
- MGF:  $M(t) = e^{t\mu+0.5\sigma^2t^2}$ .

### Exponential

- Mean:  $\frac{1}{\lambda}$ .
- Variance:  $\frac{1}{\lambda^2}$ .

### $\chi^2$

- Mean:  $r$ .
- Variance:  $2r$ .

## 26.7 Notes from Problem Sets - Part I

- Combinatorial:  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ .

- Binomial Theorem:

$$(p+q)^n = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x}.$$

- DeMorgan's:

$$\left( \bigcup_i A_i \right)^c = \bigcap_i A_i^c.$$

$$\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c.$$

- Distribution:

$$A \cap (\cup_i B_i) = \cup_i (A \cap B_i).$$

$$A \cup (\cap_i B_i) = \cap_i (A \cup B_i).$$

- Note set difference:  $A - B = A \cap B^c$ .

- Integration by Parts:

$$\int u dv = uv - \int v du.$$

- Trick:

$$e^m = \sum_{x=0}^{\infty} \frac{m^x}{x!}.$$

- $E[XY] = E[X] * E[Y] + COV[X, Y]$ .

## 26.8 Notes from Problem Sets - Part II

- Conditional Distribution:  $f(Y|X) = f(X, Y)/f(X)$ .

- $Pr(X < 4|X > 2) = Pr(2 < X < 4)/Pr(X > 2)$ .

- A and B independent:  $Pr(A \cap B) = Pr(A) \cdot P(B)$ .

- Bayes:

$$Pr(A|B) = \frac{Pr(A \cap B)}{P(B)} = \frac{Pr(B|A) \cdot Pr(A)}{Pr(B|A) \cdot Pr(A) + Pr(B|A^c) \cdot Pr(A^c)}.$$

- $V(X) = E[X^2] - (E[X])^2$ .

- Chebychev:

$$Pr(|X - \mu_x| \geq k\sigma_x) \leq \frac{1}{k^2}.$$

$$Pr(u(x) \geq \epsilon) \leq \frac{E[u(x)]}{\epsilon}.$$

- Covariance:  $\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$ .

- Correlation:  $\rho_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$ .

- Transformation technique:

$$f(Z_1, Z_2) = f(X^*, Y^*) \cdot |J|.$$