स्टेटिस्टिक्स फॉर डाटा साइंस- 1

प्रो. उषा मोहन

प्रबंधन अध्ययन विभाग

भारतीय प्रौद्योगिकी संस्थान, मद्रास

लैक्चर - 02

इंट्रोडक्शन- कोर्स ओवरव्यू

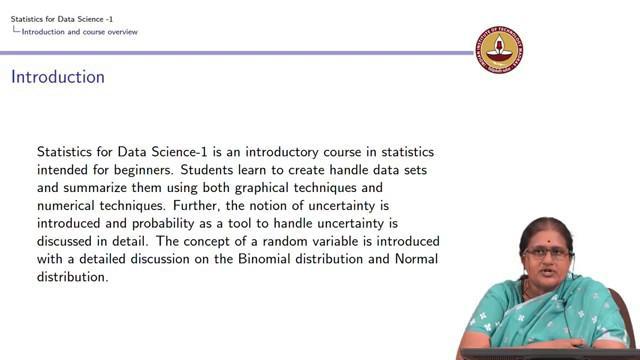
नमस्कार। यह **स्टेटिस्टिक्स फॉर डाटा साइंस-1** मॉड्यूल है। यह एक आधारभूत पाठ्यक्रम है जिसकी आपको आगे आवश्यकता होगी और यह आपको स्टेटिस्टिक्स से परिचय कराएगा। यह सांख्यिकी में एक प्रारंभिक पाठ्यक्रम है और इस लैक्चर में हम बस इस पाठ्यक्रम का अवलोकन करेंगे कि इस पाठ्यक्रम के सीखने के उद्देश्य क्या हैं और इस पाठ्यक्रम के अंत में आप क्या सीखने की अपेक्षा कर सकते हैं।

(स्लाइड समय देखें: 00:44)



इस लैक्चर में हम इस पाठ्यक्रम के संक्षिप्त परिचय के साथ शुरू करेंगे और साथ ही हम यह भी चर्चा करेंगे कि पाठ्यक्रम के सप्ताहवार कार्यक्रम और इसे सीखने के उद्देश्य क्या हैं। हम इस बात पर भी चर्चा करेंगे कि इस कोर्स कि आवश्यकता क्यों है? इस पाठ्यक्रम का मुख्य शिक्षण उद्देश्य (learning objective) क्या है?

(स्लाइड समय देखें: 01:00)

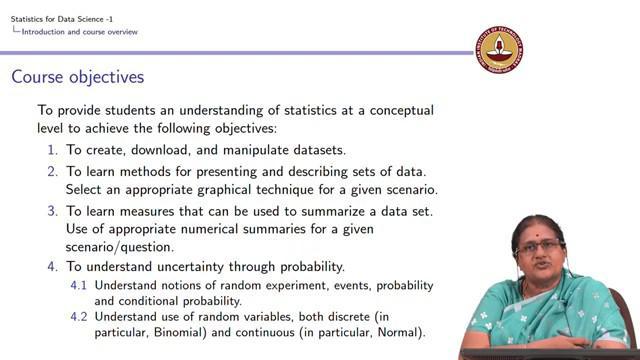


यह एक इंटरोंडक्टरी यानि परिचयात्मक पाठ्यक्रम है। यह पाठ्यक्रम बेगिन्नेर्स यानि शुरुआती लोगों के लिए है और बेगिन्नेर्स से हमारा मतलब किसी भी ऐसे व्यक्ति से है जिसने कक्षा १० के स्तर पर गणित का अध्यन्न किया हो। मुझे लगता है कि कोई भी व्यक्ति जिसने 10वीं कक्षा तक गणित पढ़ी है उसे इस कोर्स को आराम से करने में सक्षम होना चाहिए। इस पाठ्यक्रम का मुख्य आधार छात्रों को डाटा सेट बनाने मे सक्षम बनाना और डाटा को संक्षेप मे प्रस्तुत करना है। जब हम डाटा को सारांशित (summarize) करने के बारे में बात करते हैं तो हम ग्राफिकल तकनीकों और न्यूमेरिकल तकनीकों का उपयोग करने के बारे में भी बात करते हैं।

इस पाठ्यक्रम के बारे में अगली महत्वपूर्ण बात यह है कि हम इस पाठ्यक्रम मे अनिश्चितता (uncertainty) की धारणा और थ्योरी ऑफ प्रोबाबिलिटी (theory of probability) से जो समझते हैं उसे भी आपके सामने पेश करने जा रहे हैं। लेकिन हम यहाँ थ्योरी ऑफ प्रोबाबिलिटी को **यानि** अनिश्चितता की धारणा को समझने के लिए एक उपकरण (tool) के रूप में उपयोग करने पर विस्तार से चर्चा करेंगे।

अंत में हम आपको रैनडम चर (random variable) की धारणा या अवधारणा का परिचय देंगे जिसमे हम दो महत्वपूर्ण वितरणों (distribution) पर अपना ध्यान केंद्रित करेंगे; एक बाइनोमियल डिस्ट्रिब्यूशन यानि द्विपद वितरण और दूसरा नॉर्मल यानि सामान्य वितरण। फिर हम इनके बारे में अनुप्रयोगों (applications) का अध्ययन भी करेंगे। पूरे पाठ्यक्रम में हम केवल वैचारिक स्तर पर अनुप्रयोगों और अपनी समझ को विकसित करने पर अपना ध्यान केंद्रित करेंगे। इस पाठ्यक्रम मे हमारा ध्यान स्टेटिस्टिक्स के पीछे के सिद्धांत, प्रमेयों (theorem) और प्रमाणों (proofs) पर नहीं होने वाला है बल्कि हमारा फोकस क्रियान्वयन यानि एप्लिकेशन स्तर पर रहेगा।

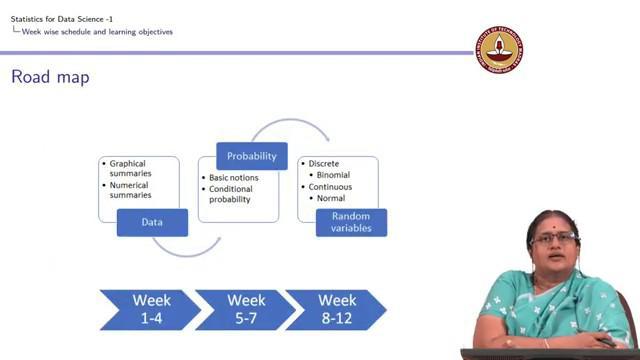
(स्लाइड समय देखें: 02:37)



तो इस पाठ्यक्रम के क्या उद्देश्य हैं? इस पाठ्यक्रम का उद्देश्य छात्रों को वैचारिक स्तर पर समझ प्रदान करना है। पाठ्यक्रम के अंत में छात्रों से यह उम्मीद की जाती है की वो समझ सकेंगे कि डाटा सेट को कैसे विकसित , डाउनलोड और कैसे कार्य के उद्देश के हिसाब से डाटा सेट मे हेरफेर की जाती है । यही इस पाठ्यक्रम से हमारे उद्देश्यों में से पहला उद्देश है। विद्यार्थी को आँकड़ों को प्रस्तुत करने और उनका वर्णन करने की विधियाँ आनी चाहिए। डाटा को प्रस्तुत करने और उसका वर्णन करने से हमारा तात्पर्य है कि हम छात्रों से यह उम्मीद करते हैं कि एक परिदृश्य को देखते हुए छात्र यह समझने में सक्षम होगा कि कौन सी उपयुक्त ग्राफिकल तकनीक का उपयोग करे। हम डाटा के ग्राफिकल सारांश और डाटा के नुमेरियाकल सारांश दोनों के बारे में बात करेंगे और बहुत जल्दी हम इस बात को समझ जाएंगे कि सभी प्रकार के डाटा के लिए न्यूमेरिकल सारांश नहीं दिए जा सकते हैं। तो एक परिदृश्य और उससे जुड़े प्रश्नो को देखते हुए हम न्यूमेरिकल सारांशों को तैयार करने पर भी ध्यान केंद्रित करेंगे।

पूरे पाठ्यक्रम के दौरान हमारा इस बात पर भी ध्यान केंद्रित रहेगा कि कैसे छात्र दिए गए डाटा से प्रश्न तैयार करेंगे। हम इस पाठ्यक्रम के दौरान अनिश्चितता को समझने के लिए प्रॉबबिलिटि का उपयोग करेंगे और अनिश्चितता को समझने के क्रम में हम रैनडम एक्सपेरिमेंट की धारणा क्या है उसे भी समझने की कोशिश करेंगे। साथ ही साथ रैनडम चर (random variable) के उपयोग को भी समझगे। तो इस पाठ्यक्रम के अंत में एक छात्र को किसी भी प्रकार के डाटा को सारांशित करने और प्रॉबबिलिटि की मूल बातें समझने का एक अच्छा वैचारिक स्तर हांसिल हो जाना चाहिए।

(स्लाइड समय देखें: 04:07)



इस पूरे पाठ्यक्रम की अवधि १२ सप्ताह की है और इस पाठ्यक्रम का रोडमैप निम्नलिखित ग्राफ में दिया गया है। तो यदि आप इस पाठ्यक्रम को मॉड्यूलराइज़ करेंगे तो हमारे पास 3 मॉड्यूल हैं। पहले मॉड्यूल में हम डाटा की मूल बातें समझेंगे; डाटा के प्रकार क्या हैं; डाटा को हम कैसे सारांशित कर सकते हैं यह भी समझेंगे। यह सब हम पाठ्यक्रम की अवधि के प्रथम 4 सप्ताह में पूरा करने की उम्मीद करते हैं।

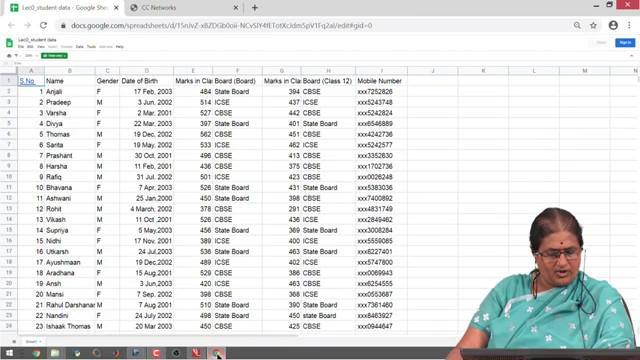
एक बार जब हम इन 4 हफ्तों में डाटा के बारे में जान लेंगे तो हमे किसी भी तरह की अनिश्चितता का सामना नहीं करना पड़ेगा। एक बार जब हम सप्ताह 4 तक के पाठ्यक्रम को पूर्ण कर लेंगे तो हम प्रॉबबिलिटि की अवधारणाओं को समझने के लिए आगे बढ़ सकते हैं जहां हम प्रॉबबिलिटि की बुनियादी धारणाओं को सीखेंगे। हम लगभग 3 सप्ताह का समय यह समझने के लिए समर्पित करेंगे कि प्रॉबबिलिटि के अनुप्रयोग क्या हैं और तत्पश्चात हम रैनडम चर (random variable) की धारणा को समझने के लिए पाठ्यक्रम मे आगे बढ़ेंगे जहां हम बिनोमिअल तथा नॉर्मल डिस्ट्रीब्यूशन (binomial and normal distribution) पर विशेष ध्यान देने के साथ साथ डिस्क्रीट और कंटिन्युस रैनडम वेरिएबल के बारे में बात करेंगे।

(स्लाइड समय देखें: 05:17)



तो आइये हम इस बात से शुरूआत करते हैं कि हमे पाठ्यक्रम से क्या अपेक्षाएँ हैं और सप्ताह दर सप्ताह अपेक्षाएं तय कर के हम पूरे पाठ्यक्रम के लिए अपेक्षाएं निर्धारित करते हैं। तो आइए हम एक बहुत ही सरल उदाहरण देखते हैं जहां मेरे पास एक विश्वविद्यालय है, मैं इसे XYZ विश्वविद्यालय नाम दे देती हूं, जिसने अभी-अभी अपने स्नातक कार्यक्रम में प्रवेश प्रक्रिया को पूरा किया है। प्रत्येक प्रवेशित छात्र से एक फॉर्म भरने के लिए कहा गया था और निम्नलिखित जानकारी सारणीबद्ध (tabulated) है।

(स्लाइड समय देखें: 05:46)



जिस क्षण मैं कहती हूं कि सूचना सारणीबद्ध है तो जैसा की आप इस डाटा को देख सकते है जो केवल काल्पनिक डाटा ही है जिसे हमने छात्रों से आवेदन फॉर्म के द्वारा जानकारी एकत्र कि थी जिसमे व्यक्ति का नाम, जेंडर, उसकी जन्म तिथि, कक्षा 10 में प्राप्त अंक, जिस बोर्ड से छात्र ने कक्षा 10 के लिए परीक्षा उत्तीर्ण की उस बोर्ड का नाम; कक्षा 12 में प्राप्त अंक और बोर्ड का नाम तथा उनका मोबाइल नंबर आदि हैं ।

पहली नज़र में आप देख सकते हैं कि हम सभी इस तरह के डाटा सेट या इस तरह के डाटा से परिचित हैं। तो, इस डाटा सेट से हम सामान्य रूप से क्या जानकारी चाहते हैं? सबसे पहले हम जो करने जा रहे हैं वह यह है कि हम इस डाटा सेट के आधार पर कुछ प्रश्न पूछना शुरू करेंगे। यह एक बहुत ही सरल डाटा सेट है और हम यह देखने की कोशिश कर रहे हैं कि पाठ्यक्रम के दौरान हम इस तरह के डाटा सेट से किस प्रकार के उत्तर पाने की उम्मीद कर सकते हैं। तो आइये देखे कि इस तरह के डाटा सेट से हम यहाँ क्या कर सकते हैं? जैसा की आप देख सकते हैं कि मैंने इस डाटा सेट के एक हिस्से को यहाँ पेस्ट किया है।

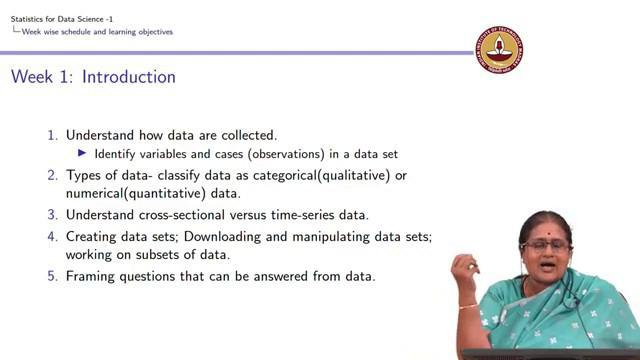
तो सबसे पहली चीज जो हम इस डाटा सेट मे पाते हैं वो हैं, जैसे की हमारे पास नाम है, हमारे पास व्यक्ति का जेंडर है। आप यह भी देख सकते हैं कि डाटा सेट मे आपके पास कक्षा १० में और कक्षा १२ में अंकों को नंबर के रूप में लिखा गया है। यहाँ हमारे पास मोबाइल नंबर भी है जो नंबर के रूप में दर्ज है।

जब मैं बोर्ड के बारे में बात करती हूं, तो हमारे पास यहाँ राज्य बोर्ड के नाम और सीबीएसई बोर्ड का नाम हैं, हमारे पास आईसीएसई बोर्ड भी यहाँ दर्ज़ है। और हमे यहाँ जन्मतिथि भी दी हुई है। तो जब भी हम इस प्रकार के डाटा सेट को देखते हैं, तो सबसे पहली चीज़ जो हम इस डाटा सेट मे पाते हैं वह यह है कि हमारे पास सिर्फ एक प्रकार का चर नहीं है। इसलिए हमें यहाँ यह समझने की जरूरत है कि वह कौन सी चीज है जो यहां भिन्नता लिए हुए है।

(स्लाइड समय देखें: 07:49)



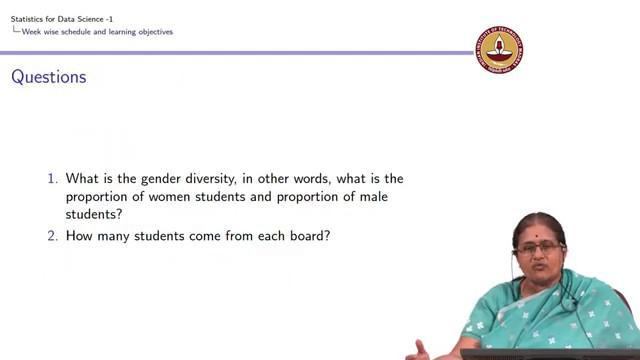
(स्लाइड समय देखें: 07:53)



इस पाठ्यक्रम के पहले सप्ताह में हम विद्यार्थियों को विभिन्न प्रकार के डाटा सेट से परिचित कराएंगे। इसी सप्ताह में हमारा पहला उद्देश्य यह समझना है कि डाटा को कैसे एकत्र किया जाता है। एक बार जब हम यह समझ जाएँगे कि डाटा को कैसे एकत्र किया जाता है तब हम, चर और अवलोकनों से क्या समझते हैं, यह जानेगे। डाटा को समझने में यह पहला कदम है। आंकड़ों के यांत्रिकी (mechanics of statistics) में छलांग लगाने से पहले हमें अपने डाटा और डाटा सेट को अच्छी तरह से समझने की जरूरत है। जब हम अपने डाटा को अच्छी तरह से समझने की कोशिश करते है तो हमें यह समझने की आवश्यकता होती है कि चर के प्रकार क्या हैं, हमने किस प्रकार का डाटा एकत्र किया है और हमारे सामने उपलब्ध विभिन्न वर्गीकरण (classification) क्या हैं। दो प्रमुख वर्गीकरण जो उपलब्ध हैं जिसमे हम डाटा को वर्गीकृत कर सकते हैं वो हैं क्वांटिटेटिव और क्वालिटेटिव डाटा यानि न्यूमेरिकल और कटेगोरिकल डाटा।

तो पहली चीज़ जिस पर हम अपना ध्यान केंद्रित करने जा रहे हैं वो है कि डाटा सेट बनाते कैसे हैं या हम कैसे मौजूदा डाटा सेट को डाउनलोड कर सकते हैं। ऐसा हो सकता है कि हम संपूर्ण डाटा सेट के साथ काम नहीं करना चाहते हों बल्कि हम केवल डाटा के सबसेट के साथ काम करना चाहते हों। तो हम सम्पूर्ण डाटा से सबसेट कैसे बनाएँगे? लेकिन यह बात ध्यान मे रखे की इस प्रथम सप्ताह मे हम यहाँ इस बात पर ज़्यादा ज़ोर देंगे की हम डाटा की मदद से कौन से प्रश्न तैयार कर सकते हैं या उन प्रश्नो के उत्तर क्या हैं जो हम डाटा से चाहते हैं। क्या हम उन सभी सवालों के जवाब अपने एकत्र किए गए आंकड़ों से दे सकते हैं जिनका उत्तर हमें पता लगाने की जरूरत है? इस स्तर पर कम से कम हम आपको डाटा से उन सवालों को तैयार करने में तथा उन सवालों के उत्तर हांसिल करने मे प्रशिक्षित करना चाहते हैं।

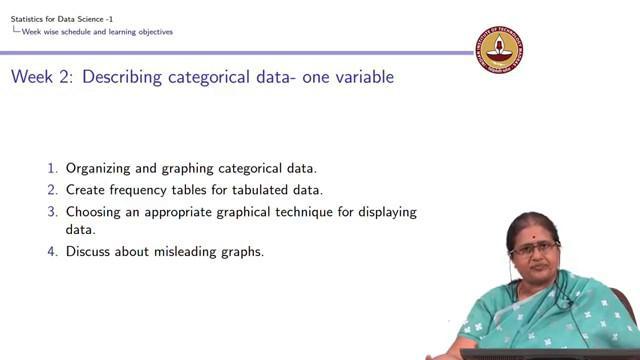
(स्लाइड समय देखें: 09:39)



एक उदाहरण के रूप में आप देख सकते हैं कि जो डाटा सेट मैंने अभी आपको प्रस्तुत किया था उससे हम क्या जानना चाहते हैं, क्या हम यह जानना चाहते हैं कि महिला छात्राओं का और पुरुष छात्रों का अनुपात क्या है। यह एक ऐसा सवाल है जो कोई भी जानना चाहेगा या आप इस डाटा से यह जानना चाहते है कि सीबीएसई (CBSE) बोर्ड से आए प्रवेशित लोगों का अनुपात क्या है? यानि सीबीएसई से आए लोगों की संख्या कुल संख्या मे कितनी है? आईसीएसई (ICSE) या विभिन्न बोर्डों से आए लोगों का वितरण क्या है? और अगर प्रवेशित लोगों का मूल राज्य की भी जानकारी एकत्र की गयी है तो क्या हम यह जानना चाहते हैं कि छात्रों का क्षेत्रीय प्रतिनिधित्व क्या है?

ऐसे कई प्रश्न हैं जिनके बारे में आप जानना चाहेंगे कि डाटा सेट मे किसी विशेष क्षेत्र के लोगों का अनुपात क्या है या किसी विशेष क्षेत्र के लोगों की संख्या कितनी है। तो इस तरह के प्रश्नो मे हमे रुचि है। इस तरह के डाटा को हम कटेगोरिकल या श्रेणीबद्ध डाटा के रूप में संदर्भित करते हैं।

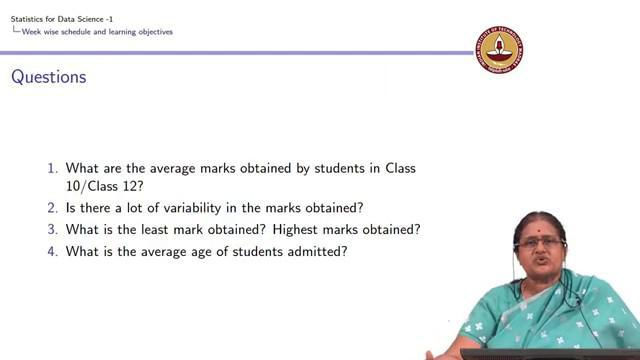
(स्लाइड समय देखें: 10:27)



इस पाठ्यक्रम के पहले तथा दूसरे सप्ताह में हम कटेगोरिकल (categorical) डाटा को समझने में अपना समय तथा ध्यान केन्द्रित करेंगे। हमारा ध्यान इन बातों पर होगा की हम कटेगोरिकल (categorical) डाटा से स्पष्ट रूप से कौन- कौन से प्रश्न जानना चाहते हैं और हमे किस प्रकार के उत्तर मिलने की उम्मीद हैं। इसी क्रम मे हम इस बात पर भी अपना ध्यान केंद्रित करेंगे कि कैसे हम फ़्रीक्वेंसी टेबल बनाकर तथा ग्राफ के उपयोग से श्रेणीबद्ध डाटा (categorical data) को व्यवस्थित कर सकते हैं। सबसे महत्वपूर्ण बात यह है कि हालांकि हमारे पास कई ग्राफिकल तकनीकें उपलब्ध हैं पर हम इस बात को समझेंगे कि किसी विशेष प्रश्न का उत्तर देने के लिए सबसे उपयुक्त ग्राफिकल तकनीक कौन सी होगी। यही वह सीखने और समझने का स्तर है जो हम आपको देना चाहते हैं जहां आप एक प्रश्न तैयार करें और पहचान कर सकें कि कौन से उपयुक्त तकनीक का इस्तेमाल करके आप उस प्रश्न का उत्तर देने सकते है।

कई बार हम पाते हैं कि बहुत सारे ग्राफ हैं जिनका इस्तेमाल डाटा को समझने मे भ्रामक हो सकता हैं। तो हम इस बात पर भी चर्चा करेंगे कि ग्राफिकल सारांशों (graphical summarizes) के माध्यम से डाटा का सही ढंग से वर्णन करना कितना महत्वपूर्ण है। जैसा की मैंने आपको पहले ही बताया की आप डाटा को कटेगोरिकल डाटा या गुणात्मक डाटा और न्यूमेरिकल डाटा यानि संख्यात्मक डाटा में व्यापक रूप से वर्गीकृत कर सकते हैं तो एक बार जब हम समझ लेंगे कि ग्राफिकल सारांशों का उपयोग करके कैसे कटेगोरिकल डाटा या गुणात्मक डाटा को सारांशित किया जाता है तो आगे हम न्यूमेरिकल डाटा को विस्तार से समझेंगे।

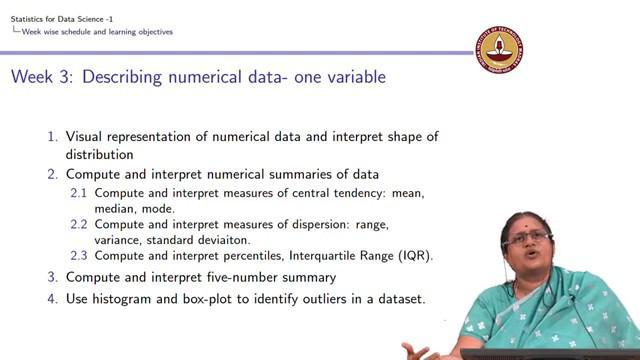
(स्लाइड समय देखें: 11:48)



तो हमारे पास यहाँ किस प्रकार के प्रश्न मौजूद हैं? जिस क्षण मैं आपसे कहती हूं कि हमारे पास जो डाटा है वो न्यूमेरिकल डाटा है तो हम समझ जाते हैं कि हम इस डाटा पर कुछ अंकगणितीय (arithmetic) ऑपरेशन कर सकते हैं और इस पर कुछ गणितीय सारांश (mathematical summarizes) प्राप्त कर सकते हैं। सबसे अधिक उपयोग किए जाने वाले सारांश में से एक को एव्रेज (average) या मीन (mean) कहा जाता है। हमको इस स्तर पर यह जानने की आवश्यकता है की ऐसे कौन से प्रश्न हैं जिनका हम इस समय उत्तर देने की आशा कर सकते हैं। अगर इस छात्रों के डाटा को हम एक बार फिर से देखें तो हम पाएंगे की किसी को भी यह जानने में रुचि हो सकती है कि छात्रों द्वारा उनकी कक्षा ११, कक्षा १० या कक्षा १२ में प्राप्त किए गए औसत अंक क्या थे।

एक और सवाल जिसको किसी को जानने की उत्सुकता हो सकती है वो यह कि क्या प्राप्त किए गए अंकों में बहुत अधिक परिवर्तनशीलता (variability) है या फिर यह की क्या लोग ऐसे अंक प्राप्त कर रहे हैं जो एक-दूसरे के बहुत करीब हैं। सबसे न्यूनतम अंक क्या है? उच्चतम अंक क्या है? छात्रों की औसत आयु क्या है? तो आप देख सकते हैं कि ये सब बहुत ही स्वाभाविक प्रश्न हैं जो हम में से किसी को भी जानने की उत्सुकता हो सकती है जब हमें इस तरह का डाटा सेट प्रस्तुत किया जाता है। दूसरे शब्दों में कहे तो हम यहां जिन प्रश्नों का उत्तर पाना चाहते हैं उनमें कुछ गणितीय सारांश या संख्यात्मक सारांश शामिल है।

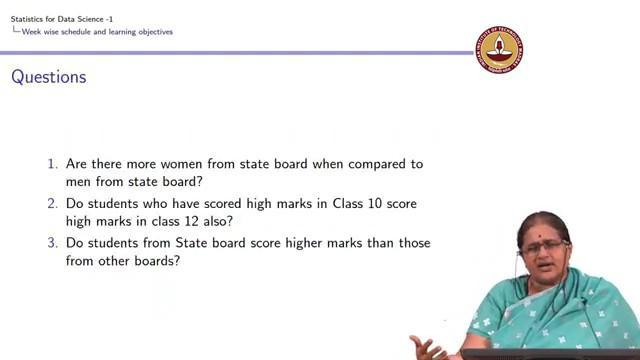
(स्लाइड समय देखें: 12:54)



तीसरे सप्ताह में हम न्यूमेरिकल सारांशों का उपयोग करके न्यूमेरिकल यानि संख्यात्मक डाटा का वर्णन कैसे करते हैं समझेंगे। साथ ही साथ हम कंटिन्युस (continuous) डाटा के ग्राफिकल सारांश पर भी चर्चा करेंगे। लेकिन इस तीसरे सप्ताह के दौरान इस बात पर अधिक ध्यान दिया जाएगा कि आप डाटा को सारांशित करने के लिए कैसे न्यूमेरिकल मापक (measure) का उपयोग कर सकते हैं। यानि मोटे तौर पर कहे तो सेंट्रल टेन्डेंसी यानि केंद्रीय प्रवृत्ति और वेरिएशन यानि भिन्नता को मापना समझेंगे।

तो यहाँ हम सेंट्रल टेन्डेंसी और वेरिएशन के मापक के बारे मे समझेंगे और साथ ही साथ हम इन्हे ग्राफिकल चीजों से भी जोड़ेंगे। मुख्य रूप से हम और आप यहाँ हिस्टोग्राम और बॉक्स प्लॉट कैसे बनाते हैं जानेगे। तो यह उम्मीद है की तीसरे सप्ताह के अंत तक आपको कटेगोरिकल और न्यूमेरिकल डाटा को कैसे सारांशित किया जाता है समझ आ जाएगा; लेकिन एक ज़रूरी बात यह की इस समय तक हम केवल एक चर के सारांश पर ध्यान केंद्रित कर रहे हैं।

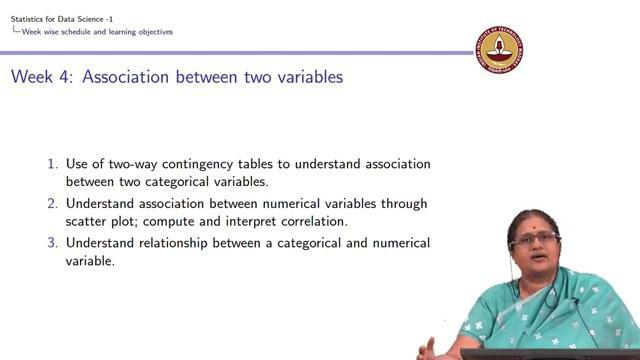
(स्लाइड समय देखें: 14:06)



आइये हम एक बार फिर से अपने स्कूल डाटा सेट पर वापस चलते हैं। डाटा सेट को देखने के बाद में आप शायद यह सवाल पूछना चाहें कि क्या महिलाएं राज्य बोर्ड से हैं, क्या राज्य बोर्ड की महिलाओं की संख्या राज्य बोर्ड के पुरुषों की तुलना में अधिक है? क्या वे लोग जिन्होंने 10वीं कक्षा में उच्च अंक प्राप्त किए हैं, उन्होने कक्षा १२ में भी उच्च अंक प्राप्त किए हैं? तो जब हम यहाँ प्रश्न पूछ रहे हैं तो हम यहाँ यह समझने की कोशिश कर रहे हैं कि क्या दो चर एक-दूसरे से संबंधित हैं या एक-दूसरे से जुड़े हुए हैं।

मैं यहां आपको सावधान कर दूँ के हम यहाँ यह नहीं पूछ रहे हैं कि क्या ए का कारण बी हैं या बी बात का कारण ए है। हम केवल इस प्रश्न का उत्तर देने में रुचि रखते हैं कि क्या 10 वीं कक्षा में अंक प्राप्त करने वाला छात्र कक्षा 12 में समान रूप से अच्छा प्रदर्शन करता है। यह एक बहुत ही स्वाभाविक प्रश्न है जैसे कि क्या राज्य बोर्ड के लोग अन्य बोर्ड के लोगों की तुलना में अधिक अंक प्राप्त करते हैं। इसलिए हम यहाँ असल में यह देख रहे हैं कि क्या हम ऐसे मापक या सारांशों का पता कर सकते हैं जो वास्तव में चर के बीच संबंधों को संक्षेप में प्रस्तुत कर सकते हैं। तो इस तरह हम एक चर को सारांशित करने से आगे बढ़ते हुए चर के बीच संबंधों को स्थापित करने की कोशिश करेंगे।

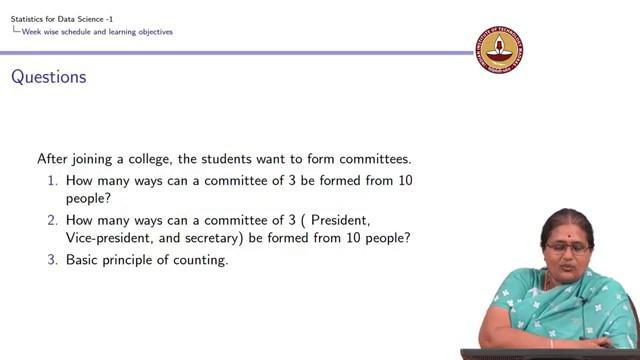
(स्लाइड समय देखें: 15:25)



तो इन्ही बातों को सीखने और समझने पर हमारा ध्यान केन्द्रित रहेगा और जो प्रश्न हम सप्ताह चार (4) में पूछने जा रहे हैं वो प्रश्न हमे दो चर के बीच संबंध के बारे में अधिक जानने मे मदद करेंगे। इसके लिए हम आकस्मिक सारणी (contingency table) और एक ग्राफिकल विधि जिसे स्कैटर प्लॉट (scatter plot) कहा जाता है का उपयोग करेगे। ऐसा कर के हम न्यूमेरिकल वेरिएबल के बारे में अधिक गहन चर्चा करेंगे और हम यह भी समझेंगे कि एक कटेगोरिकल और न्यूमेरिकल वेरिएबल एक दूसरे से कैसे संबंधित हैं। यही इस पाठ्यक्रम के पहले मॉड्यूल में शामिल है जो कि हम सप्ताह १ से ४ मे पूरा करेंगे जहां आपको एक उचित और वैचारिक स्तर का ज्ञान हांसिल होगा और आप डाटा सेट क्या है इसकी भी समझ हांसिल कर लेंगे। इसके साथ साथ आप यह भी समझ पाएंगे की आपके डाटा सेट में किस प्रकार के चर हैं और उन्हें आप कैसे वर्गीकृत कर सकते हैं; आप कैसे उन्हें मात्रात्मक, गुणात्मक या संख्यात्मक या श्रेणीबद्ध रूप में वर्गीकृत और सारांशित कर सकते हैं।

जब मैं सारांशों या समरी के बारे में बात करती हूं चाहे वह एक ग्राफिकल सारांश हो या एक न्यूमेरिकल सारांश हो तो यह बहुत महत्वपूर्ण है कि हम फिर से एक प्रश्न पूछें कि हम क्या चाहते हैं और उचित उपाय क्या है। मुझे लगता है कि यहां बहुत स्पष्ट रूप से ध्यान इस बात पर होना चाहिए कि आप किस प्रश्न का उत्तर देना चाहते हैं यानि आप अपने डाटा सेट से क्या जानकारी मांग रहे हैं और आप इसे कैसे प्राप्त करेंगे। तो यह पहला मॉड्यूल है और चौथे सप्ताह के अंत में इस मॉड्यूल को पूर्ण करने के बाद आपसे यही सब जानने की अपेक्षा की जाएगी।

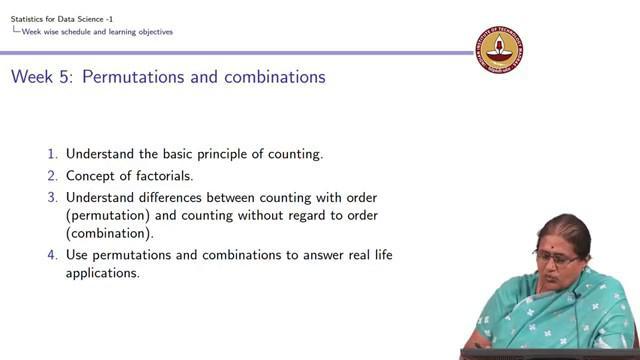
(स्लाइड समय देखें: 17:01)



दूसरा मॉड्यूल महत्वपूर्ण मॉड्यूल है जहां हम आपके सामने प्रॉबबिलिटि की धारणा को पेश करेंगे। हमें प्रॉबबिलिटि को समझने की आवश्यकता क्यों है? प्रॉबबिलिटि अत्यंत महत्वपूर्ण है क्योंकि हम अनिश्चित समय में रहते हैं और जो भी हम जानना चाहते हैं, जहां भी अनिश्चितता या एलिमंट ऑफ चान्स मौजूद होता है तो हम हमेशा प्रश्न पूछते हैं। जब भी हमारे समक्ष अनिश्चितता होती है तो हमें अनिश्चितता की इस धारणा को संभालने के लिए एक बहुत ही मजबूत टूल (tool) की आवश्यकता होती है और इस अनिश्चितता को संभालने के लिए प्रॉबबिलिटि एक अच्छा टूल है।

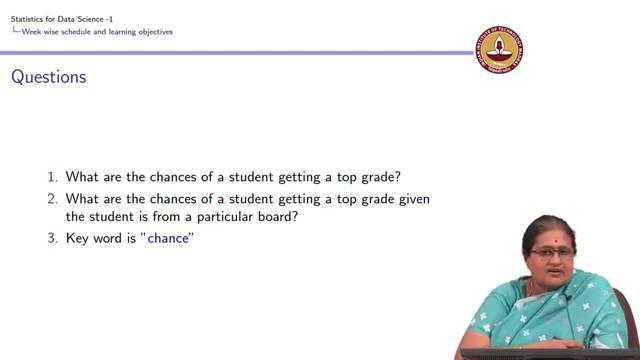
लेकिन इससे पहले कि हम समझें कि प्रॉबबिलिटि एक टूल के समान कैसे उपयोगी है हमें गणना के मूल सिद्धांत को समझने की जरूरत पड़ेगी। हमें गणना के मूल सिद्धांत (basic principle of counting) को समझने की आवश्यकता क्यों है? उदाहरण के लिए, साथ मे जुड़ने के बाद छात्र आमतौर पर एक कॉलेज में समितियां बनाते हैं। आपका सबसे स्वाभाविक प्रश्न होगा कि समिति का गठन कितने तरीकों से किया जा सकता है? तीन (3) सदस्य की समिति कितने तरीकों से बनाई जा सकती है? पूछे गए दोनो प्रश्नों के बीच का अंतर यह है कि पहले प्रश्न में हम बस यह जानने में दिलचस्पी रखते हैं कि कितने तरीकों से तीन (3) सदस्य की एक समिति बनाई जा सकती है; जबकि, दूसरे प्रश्न में हमे एक ऑर्डर जानने में दिलचस्पी है जहां हमे अध्यक्ष, उपाध्यक्ष और सचिव आदि के बारे मे जानने में दिलचस्पी है। हम में से बहुत से लोग पहले ही हाई स्कूल में इस अवधारणा से परिचित हो चुके होंगे, जिसे परम्युटेशन और कॉम्बिनेशन (permutation and combination) के रूप में जाना जाता है।

(स्लाइड समय देखें: 18:33)



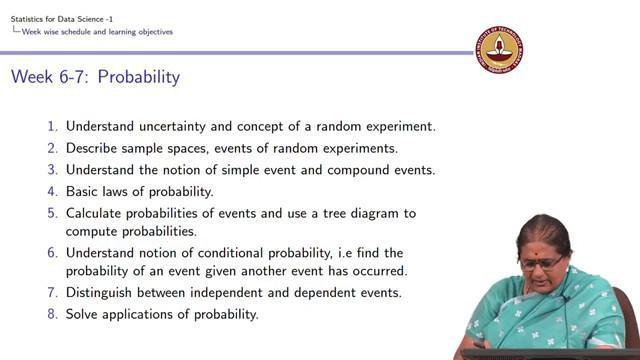
सप्ताह ५ में हम छात्रों को गणना के बुनियादी सिद्धांतों से परिचित कराएंगे और हम यह भी समझेंगे कि परम्युटेशन और कॉम्बिनेशन (permutation and combination) की धारणा को कैसे लागू किया जाता है जो कि मूल रूप से क्रम के साथ गणना और क्रम की परवाह किए बिना गणना ही है। गणना के बुनियादी सिद्धांतों का मुख्य उद्देश्य छात्रों को यह समझने में मदद करना है कि वास्तविक जीवन के अनुप्रयोगों का उत्तर देने के लिए इन परम्युटेशन और कॉम्बिनेशन का उपयोग कैसे किया जा सकता है। सप्ताह ५ के अंत में जब आप गणना के बुनियादी सिद्धांतों को कैसे लागू किया जाता है समझ जाएँगे तो उसके पश्चात, प्रश्नो को कैसे तैयार किया जाता है इस पर हम चर्चा करेंगे।

(स्लाइड समय देखें: 19:12)



एक अनिश्चितता की कल्पना कीजिए जैसे की मानिए की एक छात्र ने 10 और 12 वीं कक्षा में प्राप्त अंकों के आधार पर कॉलेज में प्रवेश लिया है। तो तत्काल वे जानना चाहेंगे कि मेरे शीर्ष ग्रेड प्राप्त करने की क्या संभावना है? अब एक प्रश्न जो प्रशासन की और से हो सकता है वो यह की किसी छात्र के टॉपर होने की क्या संभावना है इस तथ्य पर कि वे किसी विशेष क्षेत्र से आते हैं या वे किसी विशेष जेंडर से संबंधित हैं या ऐसे ही बहुत सारे प्रश्न हो सकते हैं जो आप पूछना चाहते हैं। पूछे गए सभी प्रश्न मे संभावना का अंश है यानि चान्स को समायोजित करता है, की इस बात की क्या संभावना है।

(स्लाइड समय देखें: 20:05)

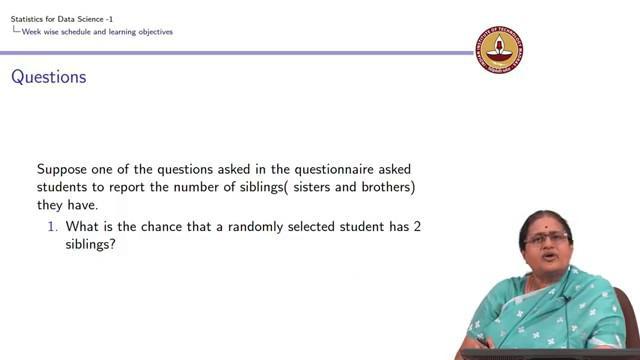


हम सब जीवन मे संभावना (chance) की इस धारणा के अभ्यस्त हैं और अगले दो (2) हफ्तों में हम छात्रों को इसी अनिश्चितता की मूल धारणा से परिचित कराएंगे। हम सभी जानते हैं की अगर आप एक सिक्का उछालते हैं तो नतीजा या तो हैड होगा या फिर टेल। लेकिन हम वास्तव में यह नहीं जानते हैं कि परिणाम हैड आयेगा या टेल। इसलिए हम यह कह सकते हैं कि इसके साथ कुछ अनिश्चितता जुड़ी हुई है। हम यहां आपसे रन्डोमनेस (randomness) की इसी धारणा का परिचय कराएंगे और इन दो (2) हफ्तों में हम यह समझने पर ध्यान केंद्रित करने जा रहे हैं कि एक साधारण इवैंट (simple event) या एक कम्पाउण्ड इवैंट (compound event) क्या है।

हम उम्मीद करते हैं की इस समय तक आते आते आप अपने गणित -1 पाठ्यक्रम में सेट के बारे में सीख लेंगे। क्योंकि संभावना और सरल घटनाओं और यौगिक घटनाओं की धारणा को समझने के लिए आवश्यकता होगी कि हम घटनाओं को सेट के रूप में प्रस्तुत कर पाएँ। घटनाओं या सेट को पेश करने के लिए हमे सेट बीजगणित (set algebra) के बारे मे ज्ञान की आवश्यकता होगी। जब तक हम पाठ्यक्रम के इस स्तर तक आएंगे आपके पास पहले से ही सेट एल्जिब्रा का ज्ञान अपने गणित -1 पाठ्यक्रम से हो चुका होगा जिसे आप यहाँ प्रॉबबिलिटि की धारणाओं को समझने और विकसित करने के लिए लागू करेंगे।

सप्ताह सात (7) के अंत तक आपको परस्पर अनन्य घटनाएं यानि mutually exclusive इवैंट क्या हैं का ज्ञान हो जाना चाहिए? स्वतंत्र घटनाओं (independent event) से हमारा क्या तात्पर्य है, अगर हम एक सिक्का उछालते हैं और मुझे हैड मिलता है- तो अगर मैं एक सिक्का दो बार उछाल रही हूं - पहले टॉस में हैड प्राप्त हुआ तो क्या यह दूसरे टॉस के मेरे परिणाम को प्रभावित करेगा या दोनों इवैंट स्वतंत्र हैं एक दूसरे से ? ये वे धारणाएँ हैं जिनका इस्तेमाल कर के हम प्रॉबबिलिटि फ्रेमवर्क को विकसित करेंगे जिससे हम अपने सवालों के जवाब हांसिल कर सके। पाठ्यक्रम मे सातवें सप्ताह के अंत तक आप इन सब बातों को समझ चुके होंगे ऐसी आपसे उम्मीद है।

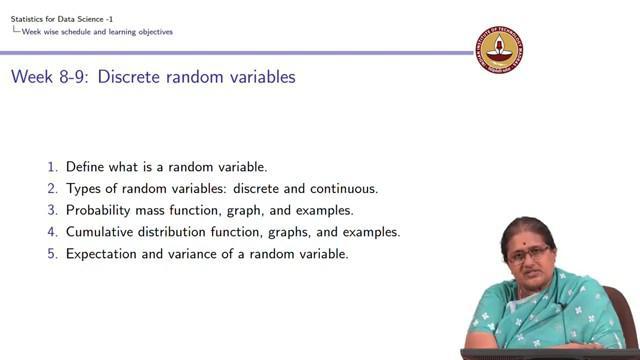
(स्लाइड समय देखें: 21:58)



अब तक हमारी बातचीत इवैंट या घटना पर केंद्रित रही। जैसा की मैंने आपसे पहले भी कहा कि हम घटनाओं के बारे में एक सेट के रूप में बात करते हैं तो हम यह भी जानते हैं कि जब भी आप एक सेट के बारे में बात करेंगे तो यह ज़रूरी नहीं है कि आपके पास हमेशा सेट के एलिमंट के रूप मे संख्याएं ही हो। लेकिन कभी-कभी हमें सवाल पूछने की जरूरत होती है जैसे की उदाहरण के तौर पर हम उसी प्रश्नावली में जिसका हम बार बार उल्लेख करते हैं -किसी छात्र को भाई-बहनों की संख्या रिकॉर्ड करने या देने के लिए भी कहा जाए की उनके कितने बहनें या कितने भाई हैं और संयोगवश आप ऐसे ही किसी छात्र का चयन करते हैं।

तो आप जिस प्रश्न का उत्तर देना चाहते हैं वह यह है कि मेरे डाटाबेस से रैनडमली चुने गए छात्र के दो (2) भाई-बहन होने की क्या संभावना है? मैं इसे केवल दो (2) तक सीमित कर रही हूं, लेकिन यह संख्या एक भी हो सकती है, यह शून्य (zero) भी हो सकती है। दूसरे शब्दों में मैं जो कुछ भी हांसिल करना चाहती हूं उसके साथ मैं एक संख्यात्मक मान (numerical value) जोड़ रही हूं और यह मैं रैनडम वेरिएबल की अवधारणा के माध्यम से करने जा रही हूं।

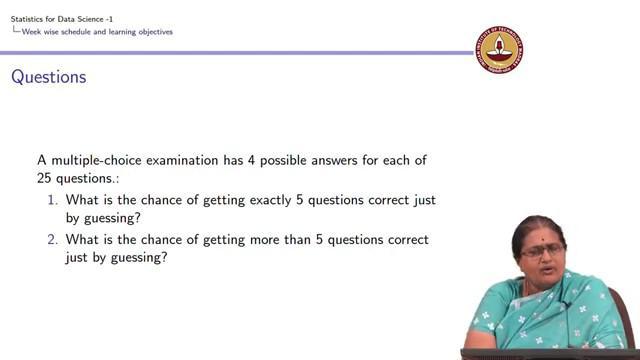
(स्लाइड समय देखें: 23:06)



हम यहाँ इस बिंदु पर रैनडम वेरिएबल की धारणा आपके समक्ष पेश कर रहे हैं और शुरुआत हम डिस्क्रीट रैनडम वेरिएबल की धारणा के साथ कर रहे जहां रैनडम वेरिएबल न्यूमेरिकल या संख्यात्मक मान लेता है। और जब हम रैनडम वेरिएबल के बारे मे समझ लेते हैं तो हम चाहेंगे की हम रैनडम वेरिएबल के द्वारा किसी प्रकार के सारांश का वर्णन करे, जिसका कोई तात्पर्य भी हो। उदाहरण के लिए आप शायद यह जानना चाहें कि विश्वविद्यालय मे औसतन कितने छात्रों के दो भाई-बहन हैं या छात्रों के भाई-बहनों की औसत संख्या क्या है।

तो ऐसे सवालों के जवाब देने की दिशा में हम अपेक्षा यानि एक्सपेकटेशन और भिन्नता यानि वेरिएशन की धारणा से आपका परिचय कराएंगे। एक बार जब हम छात्रों का परिचय डिस्क्रीट रैनडम वेरिएबल से करा देते हैं तो हम एक बहुत ही महत्वपूर्ण वितरण पर अपना ध्यान केंद्रित कर सकते हैं; हम प्रश्नों को तैयार करना सीखने मे एक (1) सप्ताह बिताएँगे।

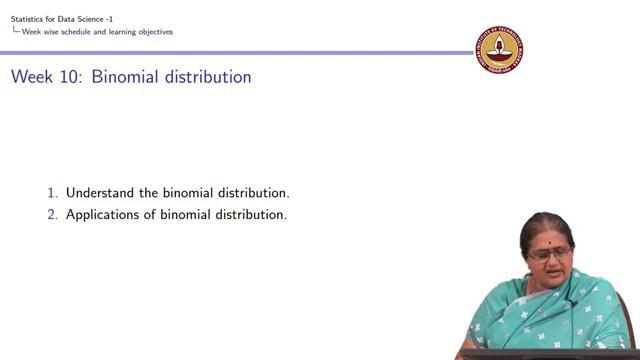
(स्लाइड समय देखें: 24:05)



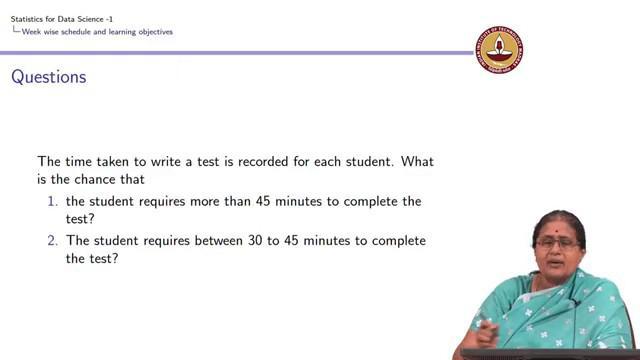
हम सभी ने कभी न कभी ऐसी परीक्षा दी होगी जिसमें बहुविकल्पीय (multiple choice) प्रश्न होते हैं। यदि आप निश्चित हैं और आप उत्तर जानते हैं तो वहां कोई अनिश्चितता नहीं है; लेकिन हम में से कई लोग बहुविकल्पीय प्रश्न में उत्तर का अनुमान लगाते हैं। तो इसलिए जब आप बहुविकल्पीय प्रश्न वाली परीक्षा लिख रहे हों, तो यह स्वाभाविक प्रश्न है कि आपके द्वारा केवल अनुमान लगाने से प्रश्नों के सही होने की क्या संभावना है। यह एक बहुत ही महत्वपूर्ण प्रश्न है। यह जानना हमारे लिए एक बहुत ही स्वाभाविक प्रश्न है।

अनुमान लगाना आखिरकार एलिमंट ऑफ चान्स ही है और यहां हम सही या गलत उत्तर प्राप्त करने पर ध्यान केंद्रित कर रहे हैं। इस प्रकार के प्रयोगों को बर्नौली प्रयोग कहा जाता है और फिर जो डिस्ट्रिब्यूशन यानि वितरण हमें हमारे उन सभी सवालों के जवाब देने में मदद करता है जैसे की प्रश्न हम अभी कर रहे हैं तो उन्हें बिनोमिअल डिस्ट्रिब्यूशन या द्विपद वितरण कहा जाता है।

(स्लाइड समय देखें: 25:10)



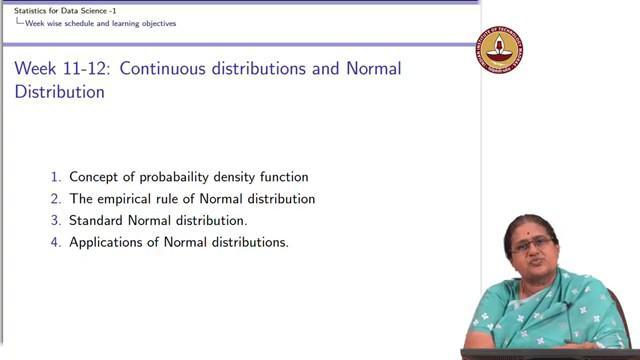
(स्लाइड समय देखें: 25:16)



इसलिए अपने पाठ्यक्रम मे सप्ताह दस (10) हम बिनोमिअल डिस्ट्रिब्यूशन को समझने मे व्यतीत करेंगे और हमारा ध्यान यहाँ फिर से बिनोमिअल डिस्ट्रिब्यूशन के अनुप्रयोगों यानि एप्लिकेशन पर केंद्रित रहने वाला है। सप्ताह १० तक हमने रैनडम वेरिएबल पर ध्यान केंद्रित किया है जो डिस्क्रीट मान लेते हैं। लेकिन कई बार हम प्रश्नों के उत्तर देने में रुचि रखते हैं। उदाहरण के लिए, यदि हम किसी छात्र द्वारा परीक्षा लिखने में लगने वाले समय को रिकॉर्ड कर रहे हैं तो हम जिन प्रश्नों का उत्तर जानने में रुचि रखते हैं वह यह है कि छात्र द्वारा परीक्षण को पूरा करने के लिए पैंतालीस (45) मिनट से अधिक समय का अनुरोध करने की क्या संभावना है या फिर यह कि छात्र को परीक्षण पूरा करने के लिए 30 से 45 मिनट की आवश्यकता होगी?

तो यहाँ आप फिर से देख सकते हैं कि जिस प्रश्न का उत्तर हम पूछ रहे हैं वह अवसर यानि चान्स के बारे में है और यहाँ हमारी रुचि का वेरिएबल समय (time) है। इसलिए हमें अपनी रुचि के इस वेरिएबल जो कि समय है को मापने का एक तरीका चाहिए और हम देख सकते हैं कि यह समय 0 मिनट, 1 मिनट, डेढ़ मिनट, एक और तीन-चौथाई मिनट के बीच कुछ भी ले सकता है। तो यह यहाँ एक कंटिन्युस वेरिएबल है। हम इस वेरिएबल को संबोधित करने पर भी अपना ध्यान केंद्रित करेंगे जिसे निरंतर चर या कंटिन्युस वेरिएबल कहा जाता है। जिसके लिए पहले हम यह पहचानेंगे कि रैनडम वेरिएबल क्या हैं जो प्रकृति में कंटिन्युस हैं।

(स्लाइड समय देखें: 26:36)



पाठ्यक्रम के अंतिम सप्ताह यानि सप्ताह ११ और १२ मे हम अपना ध्यान उन वेरिएबल पर केंद्रित करेंगे जो वास्तव में प्रकृति में कंटिन्युस वेरिएबल हैं। वहाँ हम आपको मूल रूप से प्रॉबबिलिटि डेंसिटी फंक्शन (probability density function) यानि प्रायिकता घनत्व फलन की अवधारणा का परिचय देंगे; हम मुख्य रूप से एक बहुत ही महत्वपूर्ण वितरण पर ध्यान केंद्रित करेंगे जो हमारे आँकड़ों के अध्ययन में बार-बार उत्पन्न होता रहता है जिसे हम सामान्य वितरण या नॉर्मल डिस्ट्रिब्यूशन के रूप में संदर्भित करते हैं। हम सामान्य वितरण के एम्पिरिकल रूल (empirical rule) यानि अनुभवजन्य नियम पर ध्यान केंद्रित करेंगे और फिर सामान्य वितरण के अनुप्रयोगों पर ध्यान केंद्रित किया जाएगा।

तो यह रहा हमारे पाठ्यक्रम के सप्ताह 1 से सप्ताह 12 तक का रोडमैप। सप्ताह 12 के अंत में, छात्र से अपेक्षा की जाती है कि पहले वह डाटा को समझें, डाटा सेट में हेरफेर कर सके, चर यानि वेरिएबल के प्रकारों को वर्गीकृत करना जाने। वेरिएबल के प्रकार को वर्गीकृत करने के लिए छात्र को पहले यह जानना होगा कि वेरिएबल क्या है; एक अवलोकन यानि observation क्या है और जब एक बार आप यह जान लेंगे तो आपको इन वेरिएबल को कैसे संक्षेप में प्रस्तुत करना है यह भी आना चाहिए। वेरिएबल को सारांशित करने का तरीका जानने के लिए एक छात्र से फिर से प्रश्न पूछने की अपेक्षा की जाती है कि मुझे वेरिएबल को संक्षेप में प्रस्तुत करने की आवश्यकता क्यों है; एक चर या वेरिएबल को संक्षेप में प्रस्तुत करने का उद्देश्य क्या है; मुझे किन सवालों का जवाब ढूंदना है; मैं इस डाटा सेट से क्या क्या प्रश्नो के उत्तर पा सकता हूँ? इत्यादि।

सारांश के उपयुक्त मापक क्या हैं; वेरिएबल के उपयुक्त सारांश क्या हैं? क्या हम वेरिएबलस के बीच संबंध के बारे में बात कर सकते हैं? अनिश्चितता की स्थिति में, मैं अनिश्चितता से कैसे निपटूं? हम अनिश्चित समय में रहते हैं, संभावना (probability) की मूल धारणा क्या है? रैनडम वेरिएबल की धारणा क्या है और इन रैनडम वेरिएबल के अनुप्रयोग क्या हैं? तो यह है पाठ्यक्रम का पूर्ण अवलोकन। यही अपेक्षित है। इस पाठ्यक्रम से सीखने के यही उद्देश्य हैं और बारह (12) सप्ताह के अंत में एक छात्र को अब तक जो कुछ भी प्रस्तुत किया गया है, उसकी वैचारिक स्तर की समझ के साथ सहज होना चाहिए।

धन्यवाद।

**GLOSSARY**

|  |  |  |
| --- | --- | --- |
| **ENGLISH WORD** | **WORD** | **MEANING** |
| Learning Objective | **लर्निंग औब्जैकटिव** | शिक्षण उद्देश्य |
| Theory of Probability | थ्योरी ऑफ प्रोबाबिलिटी | संभाव्यता का सिद्धांत |
| Random Variable | रैनडम वेरिएबल | अनियमित चर |
| Mechanics of Statistics | मैकेनिक्स ऑफ स्टैटिस्टिकल | आंकड़ों के यांत्रिकी |
| Probability density function | प्रॉबबिलिटि डेंसिटी फंक्शन | प्रायिकता घनत्व फलन |