स्टेटिस्टिक्स फॉर डाटा साइंस**- 1**

**प्रो. उषा मोहन**

**प्रबंधन अध्ययन विभाग**

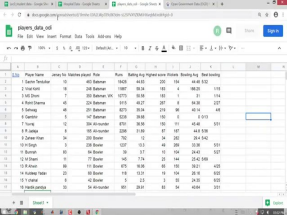
**भारतीय प्रौद्योगिकी संस्थान, मद्रास**

**लैक्चर - 05**

**इंट्रोडक्शन एंड टाइप ऑफ डाटा पार्ट- 3**

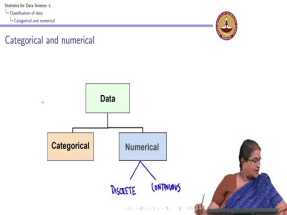
**यदि आप इस मॉड्यूल का अध्यन्न कर रहे हैं तो मैं आपसे ये अपेक्षा करती हूं कि आप यह जानते होंगे कि डाटा क्या है और आप यह भी जानते होंगे कि डाटा तालिका के रूप में डाटा को कैसे व्यवस्थित किया जाता है।**

**(स्लाइड समय देखें: 00:26)**



**और जब मैं कहती हूं कि यहाँ डाटा को डाटा तालिका के रूप में व्यवस्थित किया गया है, तो मेरा मतलब है कि आप यह भली भांति समझते हैं कि यह वह डाटा है जहां कॉलम चर का प्रतिनिधित्व कर रहे है तथा पंक्तियाँ केसेस तथा ओब्सेर्वेशन का प्रतिनिधित्व कर रही हैं।**

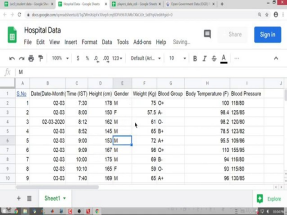
**(स्लाइड समय देखें: 00:39)**



**आइए अब एक बार हम फिर से डाटा सेट पर वापस जाते हैं। तो जैसे ही आप इस तरह के डाटा सेट को पाते हैं जैसे कि जब मैं नाम देखती हूं तो यहाँ अंजलि, प्रदीप, वर्षा, दिव्या आदि है। दूसरा चर जेंडर है। जब मैं जेंडर को देखती हूं तो मेरे पास दो श्रेणियां हैं महिला और पुरुष। जब मैं अंक को देखती हूँ तो आप भी देख सकते हैं कि यहाँ अंक दिये हुए है। मैं इस अंक को यहाँ वापस लिख देती हूँ। मैं इसे 565 के रूप में रख देती हूँ और प्रतिशत हटा देती हूँ।**

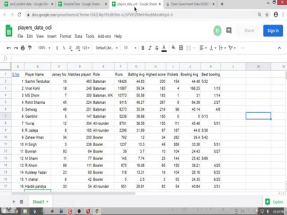
**तो जब मैं अंकों को देखती हूं, तो आप भी यहाँ देख सकते हैं कि जो अंक यहाँ पहले से मौजूद है वो हैं 484, 514, 565 वगैरह हैं, लेकिन राज्य बोर्ड फिर से यहां किसी प्रकार की श्रेणी है। मेरे पास यहाँ राज्य बोर्ड अंकित है, यहाँ आईसीएसई (ICSE) है , मेरे पास यहाँ सीबीएसई (CBSE) अंकित है। कक्षा 12 बारहवी में फिर से अंक 394, 437 तथा बोर्ड आदि अंकित हैं।**

**(स्लाइड समय देखें: 01:35)**



**आइए अब हम एक दूसरे अस्पताल के आंकड़ों पर चलते हैं। यहाँ ऊंचाई मेरे पास सेंटीमीटर मे है तो मैं इसे यहाँ सेंटीमीटर में लिखूँगी। जैसे मैं यहाँ 152, 167 और यहां 175 सेंटीमीटर लिख देती हूँ, जेंडर यहाँ फिर से पुरुष या महिला हुआ, वजन आप यहाँ देख सकते हैं 75, 57.5, 65, 98 अंकित हैं। हमारे पास यहाँ रक्त समूह, शरीर का तापमान डिग्री फ़ारेनहाइट में और रक्तचाप आदि श्रेणी है।**

**(स्लाइड समय देखें: 02:12)**



**तो जिस क्षण आप इस तरह के डाटा को देखते हैं, जैसे आइए हम क्रिकेट के आंकड़ों को देखें, तो आपके पास गेंदबाजी के औसत का डाटा है जो की 44.48, 166.25, 64.38 अंकित है जबकि खेले गए मैचों में आपके पास 463, 248, 350 है। अब, यहां आपके पास जर्सी नंबर दिये गए है जो की 10, 18, 7 वगैरह है। तो जब आप इस तरह के डाटा सेट को देखते हैं तो आपको दो चीजें उभर कर दिखती हैं। पहली चीज जो आप देखते हैं वह यह है कि आपके पास नंबर हैं और आपके पास टेक्स्ट है। टेक्स्ट से मेरा मतलब है कि आपके पास नाम हैं, आपके पास एक बल्लेबाज की भूमिका है जैसे कि ऑलराउंडर। अपने हॉस्पिटल के डाटा मे हमारे पास ब्लड ग्रुप में ओ प्लस, ए माइनस है, हमारे पास डाटा को जेंडर के तौर पर पुरुष और महिला के रूप में कैप्चर किया गया है, स्कूल के डाटा मे बोर्ड है जो सीबीएसई राज्य बोर्ड के रूप में कैप्चर किया गया है इत्यादि।**

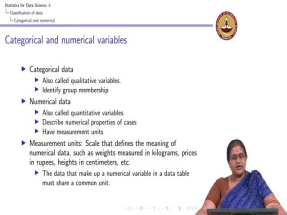
**तो आप यहाँ तुरंत देख सकते हैं कि सभी डाटा एक ही प्रकार के नहीं होते हैं। आप देख सकते हैं कि डाटा को प्रस्तुत करने के तरीके में बुनियादी अंतर है। तो अब अगली बात यह है कि क्या मैं इस डाटा को मोटे तौर पर दो श्रेणियों में वर्गीकृत कर सकती हूं।**

**हम यहाँ आसानी से देख सकते कि कुछ डाटा प्रकृति में संख्यात्मक या प्रकृति में मात्रात्मक है। उदाहरण के लिए, मैं देखती हूं कि अंक और गेंदबाजी के औसत, ऊंचाई, वजन, उच्चतम स्कोर, बल्लेबाजी का औसत, खेले गए मैच कि संख्या इत्यादि को किसी एक प्रकार के चर में क्लब किया जा सकता है जबकि नाम जेंडर, बोर्ड का नाम, रक्त समूह और खिलाड़ी कि भूमिका ये कुछ विभिन्न प्रकार के चर का प्रतिनिधित्व करते हैं।**

**एक दिलचस्प बात यह है कि अगर आप इस जर्सी नंबर को देखें, तो यह नंबर है लेकिन हम सभी जानते हैं कि इन नंबरों का कोई मतलब नहीं होता है। तो कुछ ऐसे चर हैं जो संख्यात्मक मान ले सकते हैं जैसे कि इस उधारण मे हमने पाया कि जर्सी नंबर में संख्याएँ हो सकती हैं, लेकिन वे असल मे संख्या नहीं हैं।**

**तो इसलिए, चर, जर्सी संख्या और खेले जाने वाले मैचों में निश्चित रूप से अंतर है, भले ही दोनों संख्याएं हों। इसलिए हमारे लिए यह समझना बहुत महत्वपूर्ण है कि डाटा को कैसे वर्गीकृत किया जाए और हमारे पास मौजूद चर को किस श्रेणी मे वर्गीकृत करना सही होगा। वाकई मे यह बहुत महत्वपूर्ण है। तो जो भी डाटा हम देखते हैं तो उस डाटा को मोटे तौर पर दो श्रेणियों में वर्गीकृत किया जाता है; श्रेणीबद्ध डेटा और संख्यात्मक डेटा।**

**(स्लाइड समय देखें: 05:17)**



**जब हम श्रेणीबद्ध डाटा को देखते हैं तो इन्हें गुणात्मक चर भी कहा जाता है। यह समूह सदस्यता की पहचान कराता है। समूह सदस्यता से हमारा क्या तात्पर्य है? इसको समझने के लिए हम एक बार फिर से अपने छात्र से संबन्धित डाटा पर वापस जाते हैं जहां एक चर जेंडर था। जेंडर एक श्रेणीगत चर है। मेरे पास यहां दो श्रेणियां हैं। मैं किसी भी अवलोकन को इन दो श्रेणियों में से किसी एक में वर्गीकृत कर सकती हूं।**

**तो यह एक समूह सदस्यता है। इसी तरह जब मैं स्कूल के बोर्ड के डाटा को देखती हूं तो मेरे पास एक श्रेणी है जो एक राज्य बोर्ड है, मेरे पास आईसीएसई है, मेरे पास सीबीएसई है। तो यहाँ पर भी आप देख सकते हैं कि इस श्रेणीबद्ध चर की तीन श्रेणियां हैं और किसी भी अवलोकन को इन तीन समूहों में से एक में वर्गीकृत किया जा सकता है।**

**जब हम इसे एक बार पुनः देखते हैं तो आप इस बात को आसानी से समझ सकते हैं कि मैं उस विशेष चर में एक विशेष समूह को एक अवलोकन की सदस्यता दे रही हूँ। तो इस श्रेणी में समूह हैं। आइए अब हम अस्पताल के आंकड़ों पर वापस चलते हैं। आप देख सकते हैं कि प्रत्येक रोगी का ब्लड ग्रुप या तो ओ पॉजिटिव या ओ नेगेटिव या बी पॉजिटिव या ए पॉजिटिव या ए नेगेटिव होता है।**

**तो इस तरह आप देख सकते हैं कि कई रक्त समूह होंगे। यह एक स्पष्ट रूप से श्रेणीगत चर है; जेंडर भी एक श्रेणीगत चर है। आपके हिसाब से मोबाइल नंबर किस प्रकार का वेरिएबल है? मैं इसे आपके लिए एक अभ्यास के रूप में छोड़ देती हूं। मैं चाहती हूं कि आप लोग उत्तर दें; मोबाइल नंबर आखिरकार नंबर हैं। आपके अनुसार मोबाइल नंबर किस प्रकार का वैरिएबल है?**

**इसी तरह आपको क्या लगता है कि जर्सी नंबर किस तरह का वेरिएबल है? भले ही जर्सी नंबर 10, 18, 7 हो, लेकिन यह किस तरह का वेरिएबल है? क्या यह एक श्रेणीगत चर है या यह एक संख्यात्मक चर है?**

**तो सबसे पहली बात जो हमें समझने की जरूरत है वह यह है कि मेरे पास श्रेणीबद्ध डाटा है और मेरे पास संख्यात्मक डाटा भी है। हमारे पास जो डाटा संख्यात्मक डाटा होता है उसे मात्रात्मक चर (quantitative variable) भी कहा जाता है। यहां मैं अब आपसे डाटा के संख्यात्मक गुणों (numerical properties) के बारे में बात करूंगी।**

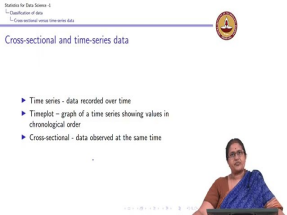
**एक बार हम फिर से पीछे जा कर अपने डाटा समहू को देखते है। कक्षा 10 और कक्षा 12 दोनों में प्राप्त अंक संख्यात्मक डाटा हैं और जब हम अंकों के बारे में बात करते हैं यह 484 अंक है, यह 514 अंक है, यह 565 अंक है। अस्पताल के आंकड़ों पर आएं तो मेरे पास 178 सेंटीमीटर, 150 सेंटीमीटर है; वजन 75 किलोग्राम, 575 क्षमा करें 57.5 किलोग्राम हैं; शरीर के तापमान के रूप मे जो डाटा डिग्री फ़ारेनहाइट के रूप मे है वो है 100 डिग्री फ़ारेनहाइट, 98.4 डिग्री फ़ारेनहाइट। जब मैं खेले गए क्रिकेट मैचों के डाटा की बात करती हूं तो मेरे पास बल्लेबाजी का औसत है, मेरे पास विकेट लेने का स्कोर है, 154 रन स्कोर है, 200 रन स्कोर है।**

**तो आप यहाँ आसानी से देख सकते है जब मैं संख्यात्मक डाटा के बारे में बात करती हूं तो या तो मैं उस डाटा के साथ माप इकाइयों को जोड़ती हूँ या मेरे पास जो डाटा है वह गेंदबाजी का औसत है या बल्लेबाजी का औसत जैसा है। आप यह भी आसानी से देख सकते हैं कि जब मैं मैचों के बारे में बात करती हूं, तो जो डाटा हमारे पास है जैसे कि 463 ये डाटा होल नंबर होता है। जब मैं बल्लेबाजी औसत के बारे में बात करती हूं तो आप देख सकते हैं कि यह कोई भी वैल्यू ले सकता है। यह फ्रक्टशन भी हो सकता है या यह कोई होल नंबर भी हो सकता है। तो जब हम श्रेणीबद्ध और संख्यात्मक डाटा के बारे में बात करते हैं तो संख्यात्मक डाटा के अंदर भी दो पहलू हैं- असतत डाटा और निरंतर डाटा। तो हम न्यूमेरिकल डाटा को और अधिक वर्गीकृत कर के कह सकते है कि न्यूमेरिकल डाटा दो प्रकार का हो सकता है एक असतत डाटा और दूसरा निरंतर डाटा।**

**तो जब हम समझ जाते हैं कि डाटा श्रेणीबद्ध डाटा है या संख्यात्मक डाटा है तब हमें उन माप इकाइयों को समझने की आवश्यकता होगी जो संख्यात्मक डाटा के लिए उपयोग की जाती हैं। आइए हम फिर से अपने डाटा पर वापस चलते हैं, आप देख सकते हैं कि यहां ऊंचाई सेंटीमीटर में मापी गयी है, वजन किलोग्राम में मापा गया है, शरीर का तापमान डिग्री फ़ारेनहाइट में दिया गया है। हमारे पास अंक हैं यह 484 अंक जो एक होल नंबर है। जब आप खिलाड़ियों के डाटा पर आते हैं तो आपके पास खेले गए मैच का डाटा हैं, उच्चतम स्कोर- यह रनों में है, विकेटों कि संख्या है।**

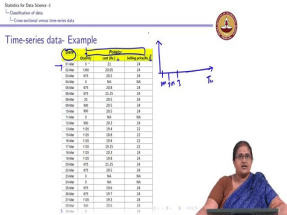
**तो यहाँ गौर करने वाली बात यह है कि हमें यह समझने की आवश्यकता है कि संख्यात्मक डाटा को परिभाषित करने वाला पैमाना क्या है। यहाँ फिर से हम इस बात पर जोर देंगे कि जब आपके पास संख्यात्मक डाटा होता है जो कि माप इकाइयाँ लेता है, तो हमें यह सुनिश्चित करने की आवश्यकता है कि चर सभी अवलोकनों के माप मे एक सामान्य इकाई साझा करे। इस बात को हमें सुनिश्चित करने की आवश्यकता है।**

**(स्लाइड समय देखें: 11:15)**



**श्रेणीबद्ध डाटा और संख्यात्मक डाटा के अलावा हमारे पास एक डाटा प्रकार और भी होता है जिसे समय श्रृंखला डाटा के रूप में संदर्भित किया जाता है।**

**(स्लाइड समय देखें: 11:27)**

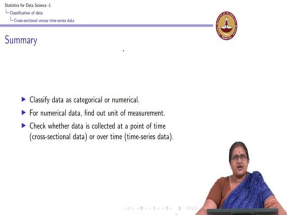


**आइए अब हम इस तरह के डाटा को देखें। आप देख सकते हैं कि इस डाटा में एक वेरिएबल है जो दिनांक है। यह डाटा वास्तव में वह डाटा है जो एक रिटेल आउटलेट पर यह ट्रैक करता है कि, किसी महीने मे, जैसे कि 1 मार्च से 30 मार्च तक हर दिन कितनी मात्रा में एक वस्तु को खरीदा जाता है साथ ही साथ खरीद की लागत और जिस कीमत पर इसे बेचा गया उसको भी दर्शाता है।**

**यदि आप इस डाटा को देखें तो यहाँ चर केवल एक चीज है जो कि आलू है। आप सभी दिनों मे ट्रैक कर रहे हैं कि आलू कितनी मात्रा में बेचा जा रहा है। दूसरे शब्दों में यहाँ हमारे पास एक तिथि है; जैसे कि पहली मार्च है, दूसरी मार्च है, मेरे पास तीसरी मार्च है। मेरे पास यहाँ समय दिया गया है और उस समय के दौरान मैं वास्तव में यह पता लगा सकती हूं कि हर दिन कितनी मात्रा मे रीटेल आउटलेट से आलू खरीदा जा रहा है।**

**तो इस तरह के डाटा को हम एक समय श्रृंखला डाटा के रूप में संदर्भित करते हैं जहां एक विशेष चर पर डाटा एकत्र होता है। यहाँ चर क्या है? यह आलू की मात्रा है जो परिवर्तनशील है, लेकिन मैं इस चर को 1 मार्च से 30 मार्च तक की अवधि में ट्रैक कर रही हूं। तो, इस तरह के डाटा को हम एक समय श्रृंखला डाटा के रूप में संदर्भित करते हैं, जबकि क्रॉस सेक्शनल डाटा वह डाटा होता है जिसे एक ही समय में देखा जाता है।**

**(स्लाइड समय देखें: 13:24)**



**तो अन्तः हम यह समझे कि डाटा को व्यापक रूप से श्रेणीबद्ध या संख्यात्मक डाटा के रूप में वर्गीकृत किया जा सकता है। जब भी हमें डाटा सेट प्रस्तुत किया जाता है तो हमें डाटा सेट में सभी चर को श्रेणीबद्ध चर या संख्यात्मक चर के रूप में वर्गीकृत करने में सक्षम होना चाहिए। यदि डाटा सेट एक संख्यात्मक चर है तो ज्ञात कीजिए कि माप की इकाई क्या है। इसके पश्चात जांचें कि क्या माप की इकाई सभी अवलोकनों में सुसंगत है। तीसरी बात यह जांचना है कि क्या यह डाटा एक समय में एकत्र किया गया है, कि क्या यह क्रॉस सेक्शनल डाटा है या फिर डाटा समय श्रृंखला डाटा है। तो अब जब भी आपको एक डाटा सेट दिया जाये तो आपको डाटा सेट को वर्गीकृत करने में सक्षम होना चाहिए।**

.

**GLOSSARY**

|  |  |  |
| --- | --- | --- |
| **ENGLISH WORD** | **WORD** | **MEANING** |
| **Representation** | **रिप्रजेंटेशन** | **प्रतिनिधित्व** |
| **Quantitative Variable** | **क्वांटिटेटिव वेरिएबल** | **मात्रात्मक चर** |
| **Numerical properties** | **न्यूमेरिकल प्रॉपर्टीस** | **संख्यात्मक गुणों** |