心理・教育測定法
基礎編A

2017年度　心理・教育測定基礎編 A・B
心理・教育測定演習 A・B

心理・教育測定法（基礎編）A・B　　担当 前川
木曜日　7-8限：W607

4月 6日	オリエンテーション	前川
4月13日	統計量の計算・相関と回帰	
4月20日	統計的推測の基礎	
4月27日	一つの母平均に関する推測	
5月11日	二つの母平均に関する推測	
5月18日	複数の母平均に関する推測	
5月25日	分散分析法と回帰分析	
6月 1日	まとめ	
6月15日	イントロ	
6月22日	多次元尺度法による個体の分類	
6月29日	クラスター分析による個体の分類	
7月 6	回帰分析	
7月13日	因子分析	
7月21日	共分散分析と構造方程式モデル	栗山
7月27日	その他の方法	
8月 4日	まとめ	

人間行動システム統計演習A・B　　担当　前川
木曜日 9-10限 W607　　栗山

1	オリエンテーション	
2	データフレームの操作	
3	ヒストグラム、クロス集計	
4	相関と回帰	
5	平均値に関する推測	
6	分散分析1	
7	分散分析2	
8	まとめ	
1	イントロ	
2	多次元尺度法	
3	クラスター分析	栗山
4	回帰分析	
5	因子分析	
6	共分散分析と構造方程式モデル	栗山
7	その他の方法	
8	まとめ	

統計的方法

- データの記述
  - 度数分布他
  - データの位置を示す指標（点推定値）
  - データの散らばり度合いを表す指標
- 統計的推測
  - 現象のモデルとしての確率分布
  - 母集団からのサンプルとしてのデータ　標本分布
  - 母集団の位置に関する推測　　統計的推定・検定
  - 複数母集団の比較
- 複数変数間の関係　→　基礎編 B

大事なこと：二つの世界

- 見える世界
  - データ、標本 (sample)
  - 標本統計量 (statistic)
- 見えない世界
  - 母集団、母集団分布
  - 母数 (parameta)　未知の定数
- 統計学の目的は
  標本（可視）から母集団（不可視）を推測

データ

- 標本サイズ、サンプルサイズ、　N, n
  - 観測対象である個体の数、
    number of observations, # of records
- 変数 variable
  - 観測した内容のこと
- 変数の数値的性質
  - 尺度水準
  - 連続か、離散か
  - 計測値か、計数値か

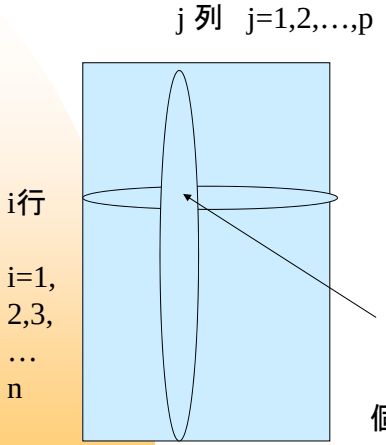
成績データ

ID	Gen	T1	Sco	GPA	ID	Gen	T1	Sco	GPA	ID	Gen	T1	Sco	GPA
1	1	165	60	3	18	1	114	56	3	35	1	155	46	3
2	1	127	62	3	19	0	140	45	4	36	1	85	66	3
3	1	146	50	4	20	1	91	70	4	37	0	221	86	4
4	1	109	47	2	21	1	93	60	3	38	1	116	45	3
5	0	72	19	3	22	1	159	44	3	39	0	118	51	4
6	0	70	40	3	23	0	141	28	2	40	0	110	40	4
7	0	88	37	3	24	1	131	70	3	41	1	123	52	3
8	0	33	37	2	25	1	143	67	4	42	1	146	46	5
9	1	196	71	5	26	0	224	90	4	43	0	89	29	2
10	1	142	72	3	27	0	126	77	3	44	1	144	69	5
11	0	170	61	4	28	0	202	63	5	45	0	91	46	3
12	1	217	77	5	29	1	93	49	2	46	0	55	0	1
13	0	106	59	2	30	0	54	9	2	47	1	52	53	3
14	1	128	42	3	31	1	108	34	2	48	1	157	57	3
15	1	32	14	1	32	1	92	32	3	49	0	93	34	3
16	0	152	57	2	33	0	133	56	3	50	0	49	37	2
17	0	39	36	1	34	0	0	11	2					

尺度水準

- 尺度の4つの種類
  - 比率尺度
  - 間隔尺度
  - 順序尺度
  - 名義尺度

データ行列



個体 i の変数 j による測定値

有名人データ

http://note.chiebukuro.yahoo.co.jp/detail/n5672

	name	height	weight	gender	genre
1	香取慎吾	180.0	71	M	歌手
2	稲垣吾郎	176.0	63	M	歌手
3	木村拓哉	176.0	74	M	歌手
4	草彅剛	170.0	60	M	歌手
5	中居正広	165.0	55	M	歌手
10	大野智	166.0	52	M	歌手
92	木村カエラ	154.0	45	F	歌手
93	米良美一	142.7	35	F	歌手
94	中村昌也	192.0	80	M	俳優
95	阿部寛	189.0	75	M	俳優
96	城田優	188.0	96	M	俳優
97	松重豊	188.0	72	M	俳優
367	ザ・たっち	151.0	43	M	その他
368	池乃めだか	150.0	35	M	その他
369	猫ひろし	147.0	35	M	その他
370	山崎静代(南海)	182.0	64	F	その他
371	冨永愛	179.0	70	F	その他
372	熊井友理奈	176.0	59	F	その他

データの記述、要約

- データのグラフ（視覚化）
  - 幹葉表示 stem and leaf
  - 度数分布図、ヒストグラム histogram
  - 箱ひげ図 box and whisker plot
- データの数値的要約
  - データの中心の位置
  - データの散らばりの大きさ
  - データの形、位置

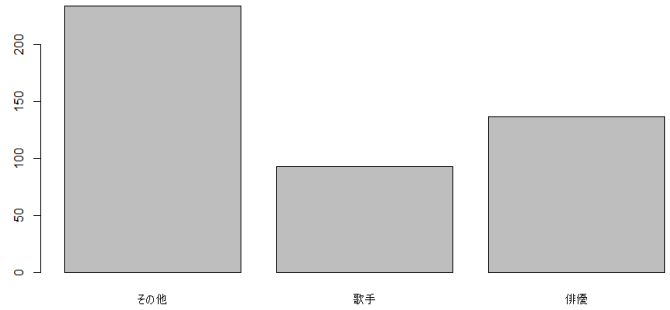
離散データの度数分布


```
>table(gender)
gender
  F   M
203 261
```

```
>table(genre)
genre
その他 歌手 俳優
   234   93  137
```

度数分布表と度数分布図

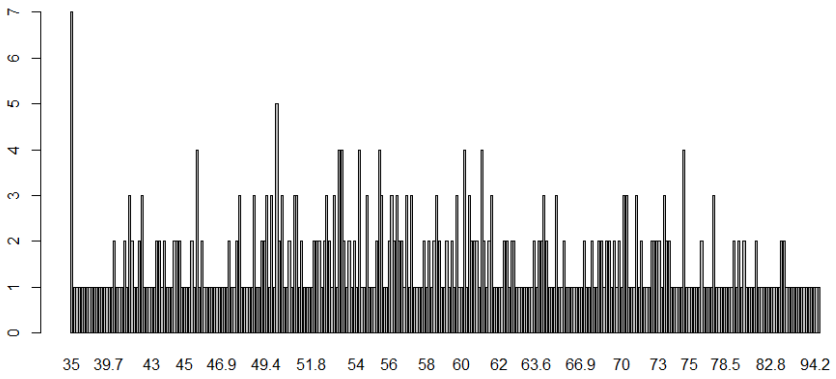

```
>table(genre)
genre
その他 歌手 俳優
   234   93  137
```



離散データの度数分布


```
> table(weight)
weight
 35 35.2 35.6 36.6   37 37.4 37.5 37.6 37.9 38.1 38.8 38.9 39.1 39.2 39.4 39.7 39.8   40 40.3 40.6 40.7 40.9   41
  7    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1    1    1    2    1

41.1 41.6   42 42.2 42.5 42.6 42.7 42.8 42.9   43 43.1 43.2 43.6 43.9   44 44.1 44.3 44.5 44.6 44.7 44.8 44.9   45
  3    2    1    1    2    3    1    1    1    1    1    2    2    1    2    1    1    1    2    2    2    1    1

45.3 45.4 45.5 45.7 45.8 45.9   46 46.1 46.2 46.3 46.4 46.5 46.6 46.8 46.9   47 47.2 47.3 47.4 47.7 47.9   48 48.1
  1    1    2    1    4    1    2    1    1    1    1    1    1    1    1    1    1    2    1    1    2    3    1

48.2 48.3 48.4 48.5 48.6 48.7 48.9 49.1 49.3 49.4 49.7 49.8   50 50.1 50.2 50.3 50.4 50.6 50.7 50.9   51 51.1 51.2
  1    1    1    1    3    1    1    2    2    3    1    3    1    5    2    3    1    1    2    1    3    3    1

51.3 51.4 51.5 51.7 51.8 51.9   52 52.3 52.5 52.6 52.7 52.8 52.9   53 53.1 53.2 53.3 53.4 53.5 53.7 53.8 53.9   54
  2    1    1    1    1    2    2    2    1    2    3    2    1    3    1    4    4    2    1    2    1    2    1

54.1 54.3 54.6 54.7 54.8 54.9   55 55.1 55.3 55.5 55.8 55.9   56 56.2 56.3 56.4 56.6 56.8 56.9   57 57.1 57.2 57.3
  4    1    1    3    1    1    1    2    4    3    1    1    2    3    2    3    2    2    1    3    1    3    1

57.4 57.5 57.6 57.9   58 58.1 58.4 58.5 58.6 58.7 58.9   59 59.1 59.2 59.3 59.5 59.6 59.8   60 60.1 60.2 60.4 60.5
  1    1    1    2    1    2    1    2    3    2    1    1    2    1    2    1    3    1    1    4    1    3    2

60.6 60.7 60.8 60.9 61.2 61.4 61.5 61.6 61.7 61.8   62 62.1 62.2 62.3 62.4 62.5 62.6 62.8 62.9   63 63.1 63.2 63.3
  2    2    1    4    2    1    2    3    1    1    1    1    2    2    1    2    2    1    1    1    1    1    1

63.4 63.5 63.6 63.7 63.8 63.9   64 64.1 64.3 64.4 64.5 64.8 65.5 65.6 65.9   66 66.1 66.2 66.5 66.8 66.9   67 67.1
  1    2    1    2    2    3    2    1    1    1    3    1    1    2    1    1    1    1    1    1    1    2    1

67.3 67.4 67.6 67.7 67.8   68 68.3 68.9 69.2 69.3 69.4 69.5 69.9   70 70.2 70.6 70.9   71 71.1 71.2 71.4 71.5 71.6
  1    2    1    1    2    2    1    2    2    1    2    1    2    1    3    3    1    1    1    3    1    2    1

71.7   72 72.2 72.7 72.8   73 73.1 73.2 73.6 73.9   74 74.1 74.4 74.5 74.7 74.8 74.9   75 75.1 75.2 75.3 75.5 75.9
  1    1    2    2    2    2    1    3    2    2    1    1    1    1    1    4    1    1    1    1    1    1    2

76.3 76.4 76.5 76.6 76.7 76.8 77.2 77.4 78.3 78.5 78.9   79 79.1 79.2 79.3 79.4 79.7 80.1 80.3 80.5 80.6 80.7 81.3
  1    1    1    1    3    1    1    1    1    1    1    1    2    1    2    1    2    1    1    1    1    2    1

81.4 82.1 82.5 82.6 82.8 83.5 83.6 83.8   84 84.2 84.9 85.1   86 86.3 86.4 86.5 86.8 87.2 87.8 88.9 92.7 92.8 94.2
  1    1    1    1    1    1    1    1    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1

96.1
   1
```

度数分布図



連続データの度数分布

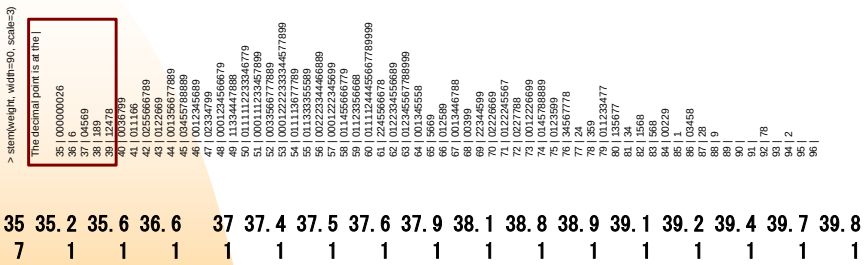
- Stem and Leaf Plot

stem(weight, width=90, scale=3)

```
> stem(weight, width=90, scale=3)

  The decimal point is at the |

  35 | 000000026
  36 | 6
  37 | 04569
  38 | 189
  39 | 12478
  40 | 0036799
  41 | 011166
  42 | 025566789
  43 | 0122668
  44 | 001356677889
  45 | 03455788889
  46 | 0012345689
  47 | 02334799
  48 | 0001234566679
  49 | 113344447888
  50 | 011111223334677 9
  51 | 0001112334557899
  52 | 0002557778899
  53 | 0001222333344577899
  54 | 011113677789
  55 | 011333355589
  56 | 0022233444668889
  57 | 0001222345699
  58 | 01145566677 9
  59 | 01223356668
  60 | 01124455667789999
  61 | 1245566678
  62 | 012234556689
  63 | 012345677889 99
  64 | 001345558
  65 | 5669
  66 | 012589
  67 | 0013446788
  68 | 00234
  69 | 2234599
  70 | 02226669
  71 | 0122245567
  72 | 0227788
  73 | 0012226699
  74 | 0145788889
  75 | 0123599
  76 | 34567778
  77 | 24
  78 | 359
  79 | 012233477
  80 | 135677
  81 | 34
  82 | 1568
  83 | 568
  84 | 00229
  85 | 1
  86 | 13458
  87 | 28
  88 | 9
  89 |
  90 |
  91 |
  92 | 78
  93 |
  94 | 2
  95 |
  96 |
```

連続データの度数分布

- Stem and Leaf Plot



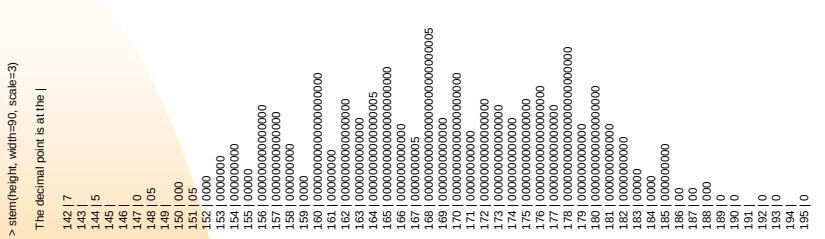
35	35.2	35.6	36.6	37	37.4	37.5	37.6	37.9	38.1	38.8	38.9	39.1	39.2	39.4	39.7	39.8
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The decimal point is at the |

35 | 000000026
36 | 6
37 | 04569
38 | 189
39 | 12478

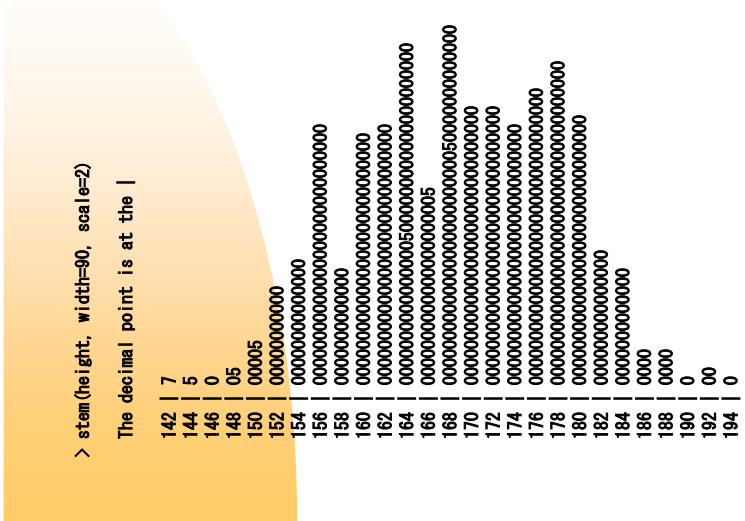
連続データの度数分布

- Stem and Leaf of Height



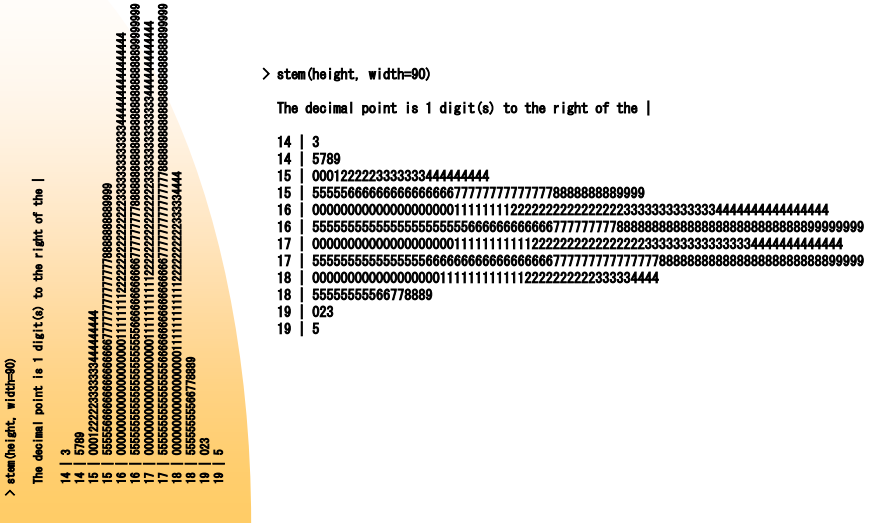
連続データの度数分布

- 離散化してから計算する



連続データの度数分布

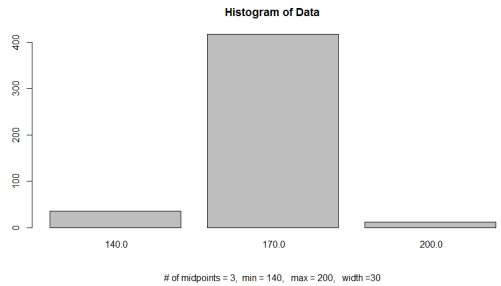
- 離散化してから計算する




```
> stem(height, width=90)

  The decimal point is 1 digit(s) to the right of the |

  14 | 3
  14 | 5789
  15 | 0001222223333333444444444
  15 | 555556666666666666677777777777777888888889999
  16 | 000000000000000000011111111122222222222222233333333333344444444444444444
  16 | 555555555555555555555566666666666677777777788888888888888888888888899999999
  17 | 000000000000000000011111111111222222222222222333333333333344444444444444
  17 | 555555555555555566666666666666677777777777777888888888888888888888899999
  18 | 000000000000000011111111111122222222233334444
  18 | 55555555566778889
  19 | 023
  19 | 5
```

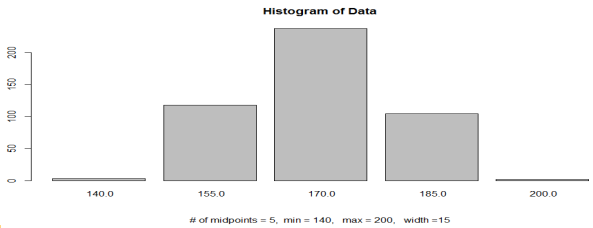
連続データの度数分布


```
Frequency Distribution

 n  =   464 ,   nobs  =   464

 npoints  =   3 ,   min  =   140 ,   max  =   200 ,   width  =   30

Categorization of Variable:  frequency
    NA  <   140.0  <=   155.0        35
 155.0  <   170.0  <=   185.0       417
 185.0  <   200.0  <=      NA        12
```



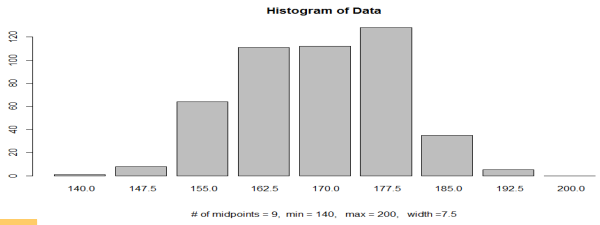
連続データの度数分布


```
Frequency Distribution

 n  =   464 ,   nobs  =   464

 npoints  =   5 ,   min  =   140 ,   max  =   200 ,   width  =   15

Categorization of Variable:  frequency
    NA  <   140.0  <=   147.5        3
 147.5  <   155.0  <=   162.5      118
 162.5  <   170.0  <=   177.5      237
 177.5  <   185.0  <=   192.5      104
 192.5  <   200.0  <=      NA        2
```



連続データの度数分布


```
Frequency Distribution

 n  =   464 ,   nobs  =   464

 npoints  =   9 ,   min  =   140 ,   max  =   200 ,   width  =   7.5

Categorization of Variable:  frequency
    NA  <   140.0  <=   143.8        1
 143.8  <   147.5  <=   151.2        8
 151.2  <   155.0  <=   158.8       64
 158.8  <   162.5  <=   166.2      111
 166.2  <   170.0  <=   173.8      112
 173.8  <   177.5  <=   181.2      128
 181.2  <   185.0  <=   188.8       35
 188.8  <   192.5  <=   196.2        5
 196.2  <   200.0  <=      NA        0
```



ヒストグラム：度数分布


```
147.5  <   155.0  <=   162.5        118
```

break 境界値	midpoint 代表値	break 境界値	frequency 度数
b_{j-1}	< x_j <=	b_j	f_j

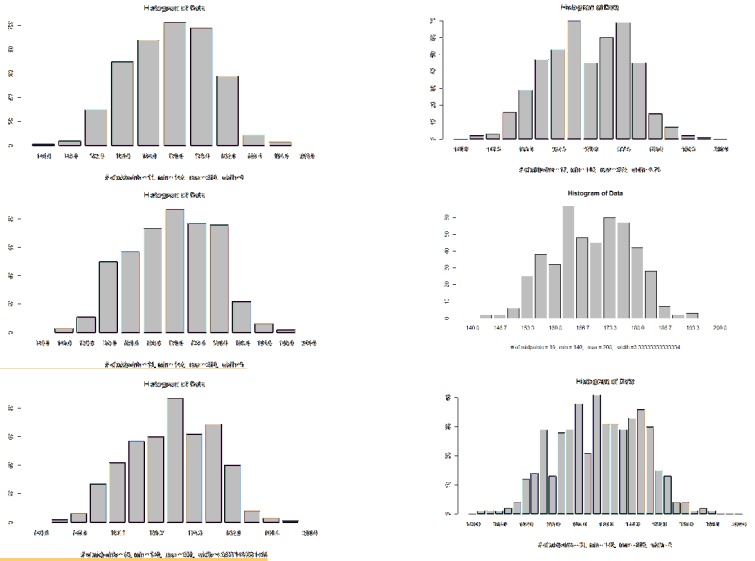
度数分布、相対度数分布

$$(x_j, f_j), \quad j = 1, 2, \ldots, k \qquad \sum_{j=1}^{k} f_j = n$$

$$(x_j, p_j), \quad j = 1, 2, \ldots, k$$

$$p_j = \frac{1}{n} f_j, \quad j = 1, 2, \ldots, k \qquad \sum_{j=1}^{k} p_j = 1$$

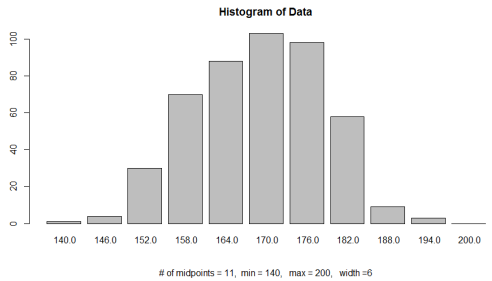
ヒストグラム



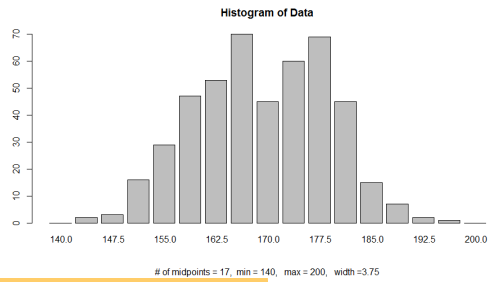
ヒストグラム

- 相対度数を用いたヒストグラムは母集団における連続的確率変数の分布を推定している。
- 標本数 n を多く取り、カテゴリ数を増やしていけば、ヒストグラムは母集団分布に近づくはず。
- 最適なカテゴリ数は？
- ちょっと冗長！？

データの中心の位置



一山（単峰）の分布



二山（双峰）の分布

データの中心の位置

- 最頻値 mode
  - 頻度が最も多いカテゴリの値
- 中央値 median
  - 全データを並べ替えた時の真ん中の値
- 平均値 mean
  - ヒストグラムを下から支えてバランスが取れる点

点の選択に伴う損失

- 中央値

$$loss = \sum_{i=1}^{n} (y_i - c)^2 \approx \sum_{j=1}^{k} f_j (x_j - c)^2$$

- 平均値

$$loss = \sum_{i=1}^{n} |y_i - c| \approx \sum_{j=1}^{k} f_j |x_j - c|$$

平均値の計算

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \approx \frac{1}{n} \sum_{j=1}^{k} f_j x_j = \sum_{j=1}^{k} p_j x_j$$

$$\sum_{j=1}^{k} f_j = n$$

$$\sum_{j=1}^{k} p_j = 1$$

分布関数

- 相対累積分布関数
  - x_k より小さいデータの割合

$$F_k = \sum_{j=1}^{k} f_j \qquad P_k = \sum_{j=1}^{k} p_j$$

$$0 \leq P_k \leq 1$$

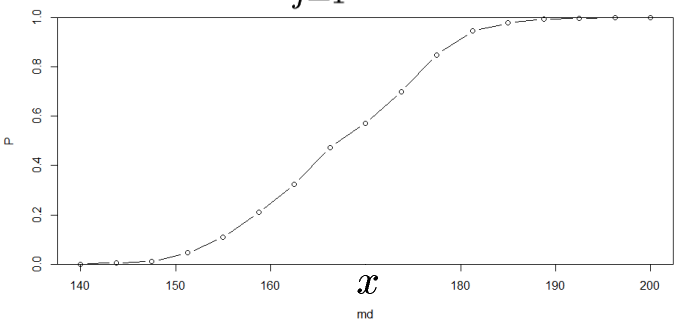
カテゴリの代表値の関数としての分布関数

$$P(x_k) = \sum_{j=1}^{k} p_j$$

分布関数

データの値（カテゴリの代表値）を与えると比率を返す関数

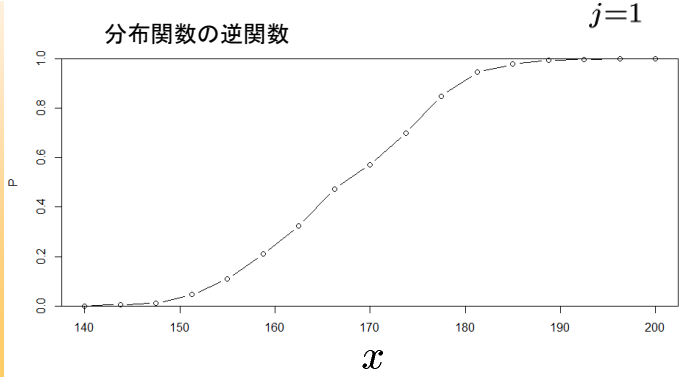
$$P(x_k) = \sum_{j=1}^{k} p_j$$



分位点

比率を与えるとデータの値を返す関数

$$x = P^{-1}(p) \qquad P(x_k) = \sum_{j=1}^{k} p_j$$

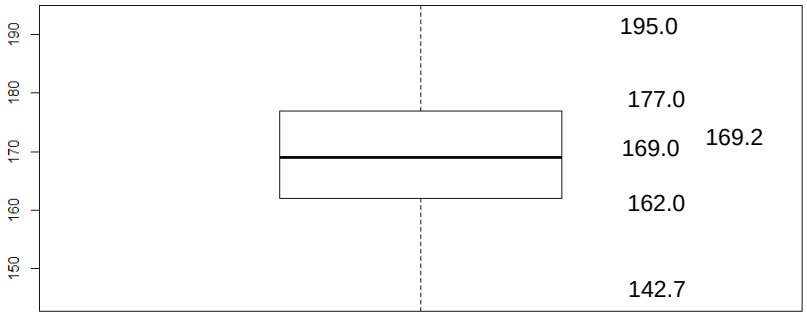


分位点

- データの分布の中で、下から
  100 x p % の位置にあるデータの値
- p の値としては
  0.10, 0.25, 0.50, 0.75, 0.90
- 四分位値 quartile

$$Q_1 = P^{-1}(0.25), \quad Q_2 = P^{-1}(0.50), \quad Q_3 = P^{-1}(0.75)$$

箱ひげ図



標準偏差 = 9.48, IQR = 15, 四分位偏差 = IQR/2 = 7.5

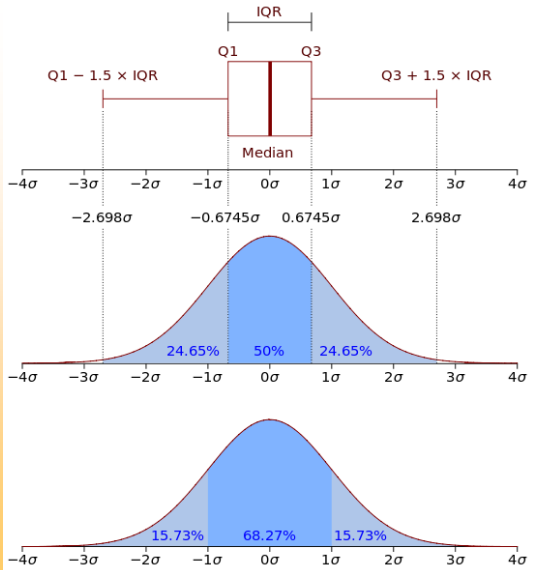
散らばり具合の指標

- 範囲 レンジ range
  - 最大値 – 最小値
- 四分位範囲 Inter Quartile Range IQR
  - Q3-Q1
- 標準偏差 = 分散 S^2 の平方根 S

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \approx \quad \sum_{j=1}^{k} p_k (x_k - \overline{y})^2$$

$var\,(y)$ variance, standard deviation, sd or std

大体の目安



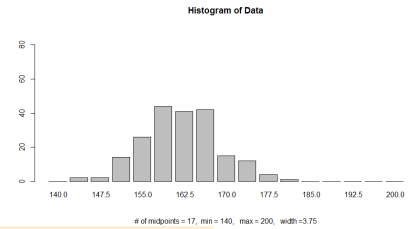
中央値±0.5 IQR = 50%

平均±1 標準偏差≈68%

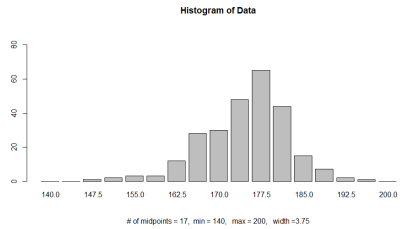
平均±2 標準偏差 ≈95%

平均±3 標準偏差≈99.7%

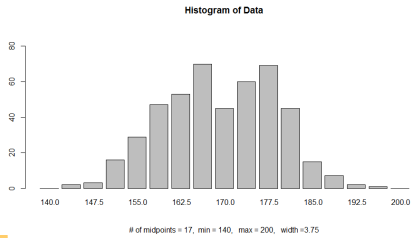
層別の分布



女性

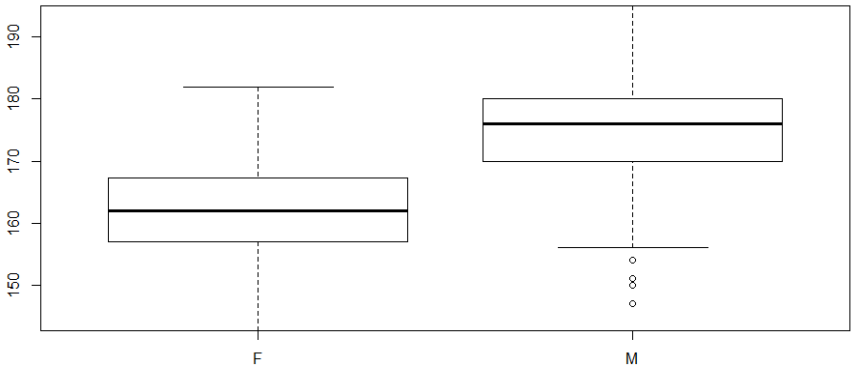


男性

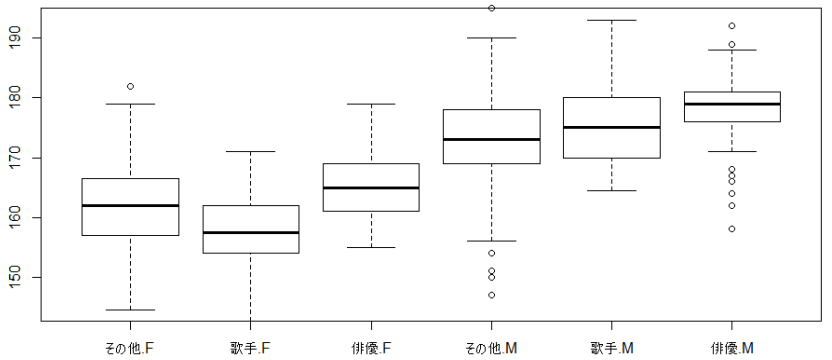


全体

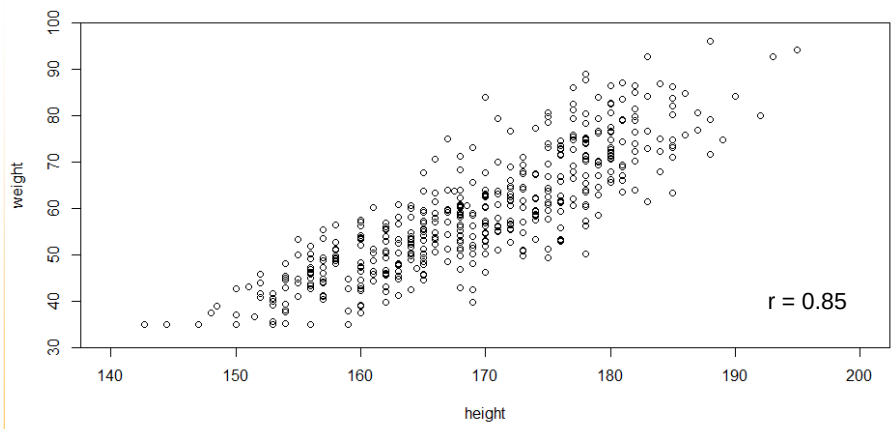
層別の分布



層別の分布



二変数のプロット：散布図



相関係数

- 共分散

$$C_{y,z} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})$$

- 相関係数

  共分散をそれぞれの標準偏差で除したもの

$$r_{y,z} = \frac{c_{y,z}}{\sqrt{var(y)var(z)}} \qquad 0 \le r_{y,z} \le 1$$

層別の散布図