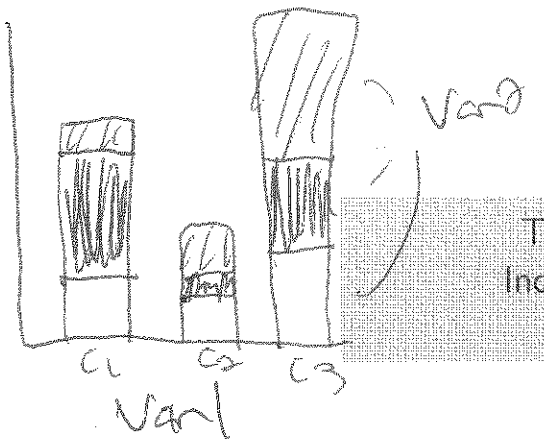
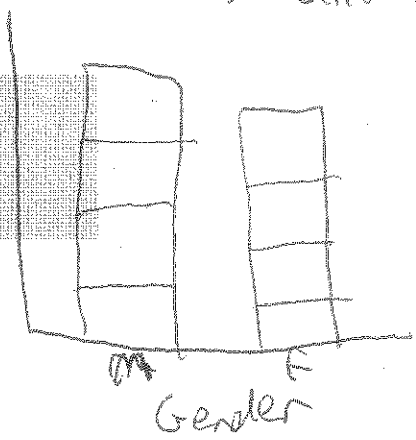


aes(X = gender, fill = class_year)
↳ marginal ↳ conditional



Today: 2-D Caterogical Data Independence and Mosaic Plots 1-D Continuous Data

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

September 19, 2016

{ } = cond. distn of Var 2 given Var1 = cat 1

2 categorical variables: Var1 + Var2

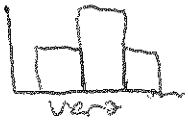
O_{ij} = # of obs. with category i in Var1 & cat j in Var2

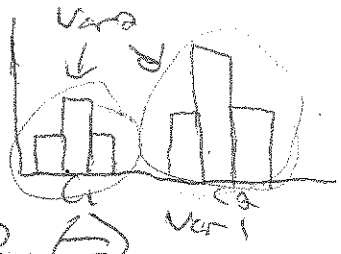
Var1: K_1 categories; Var2: K_2 categories

Contingency Tables and Marginal/Conditional Distributions

Var1 \ Var2	cat 1	cat 2	...	cat K_2	marginal distn of Var1
cat 1	O_{11}	O_{12}	...	O_{1K_2}	$n_{1\cdot}$
cat 2	O_{21}	O_{22}	...	O_{2K_2}	$n_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
cat K_1	$O_{K_1 1}$	$O_{K_1 2}$	...	$O_{K_1 K_2}$	$n_{K_1\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot K_2}$	

$n_{\cdot 1}$, $n_{\cdot 2}$, ..., $n_{\cdot K_2}$: Marginal distn of Var2





Suppose we have two events A, B

$A \perp\!\!\!\perp B \iff$

Recall: Independence Rules from Probability

$P(A) = P(A|B)$ → marginal distn of Var1 = conditional distn of Var1 | Var2

$P(B) = P(B|A)$ → -- Var2 = -- Var2 | Var1

$P(A \cap B) = P(A) \cdot P(B)$ → Joint distn of Var1 and Var2 = (marginal of Var1) × (marginal of Var2)

Can input contingency tables into chi-square tests for independence

E.g. `chisq.test(table(var1, var2))`

More on this in Lab 04 → contingency table in R

Pearson Residuals → assumes Var1 $\perp\!\!\!\perp$ Var2

→ independent

Pearson Residuals: Scaled difference between observed/expected

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{\text{observed} - \text{Expected}}{\sqrt{\text{expected}}}$$

$r_{ij} > 0$: "too many" observed } assuming Var1 and Var2 are independent

$r_{ij} < 0$: "too few" observed }

r_{ij}'s are Asymptotically Normally distributed!

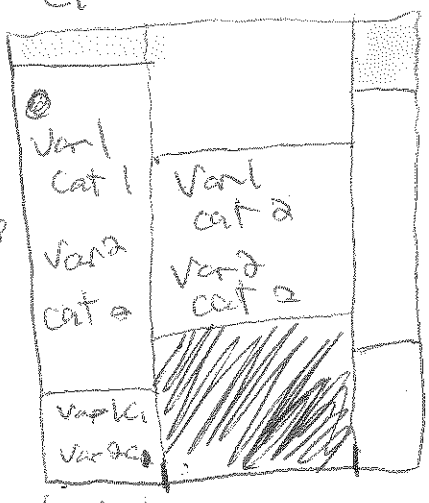
$|r_{ij}| > 2 \Rightarrow$ significant at the $\alpha = 0.05$ level

$|r_{ij}| > 4 \Rightarrow$ -- -- -- -- -- -- -- $\alpha = 0.0001$ level

color Mosaic plot according to the Pearson residuals in each cell

Mosaic Plots → visualizes contingency tables

Mosaic Plots: Area plot for two categorical variables



each cell in the contingency table gets a box in the plot

area of each box ∝ % of obs. in the corresponding cell of the cont. table

width of box i ∝ % of obs in Var1 cat i
↳ marginal distribution of Var1

height of box j ∝ % of obs. in Var2 cat j given Var1 is in cat i → conditional distribution

Can color the boxes by their differences from what was expected

Friday: Mosaic Plots in R

1-D Continuous Data

Structure: $X = \{X_1, X_2, \ldots, X_n\}$, $X_i \in \mathbb{R}$ (−∞ to ∞)
nx1 vector (column of our data), each obs is a real #

Ex: time, age, temperature, distance, height, weight, rate, probabilities

Summary:

mean / average, median (or other measures of center)

spread: standard deviation, variance, IQR

range: min, max, quantile, percentiles

In R:

range(), min(), max()

mean(), sd(), median()

summary(), var(), quantile()

modality → unimodal, bimodal, multimodal

1-D Continuous Distributions

symmetry

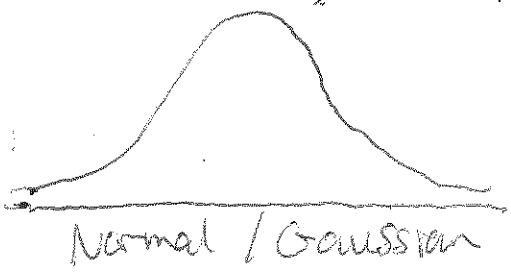
How do we describe continuous distributions?

outliers

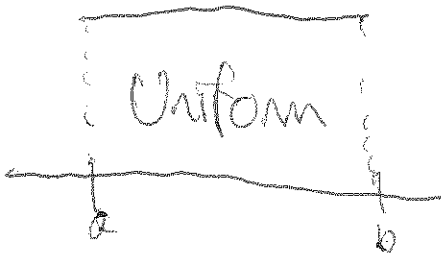
- Shape (skew)
- Center (mean, median)
- Spread (sd, var)

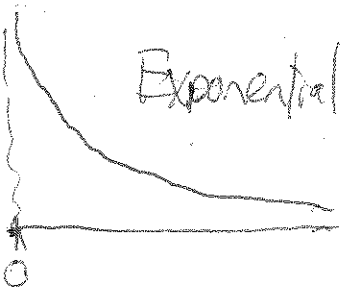
ranges of values that are most common → "high density"
or least common → "low density"

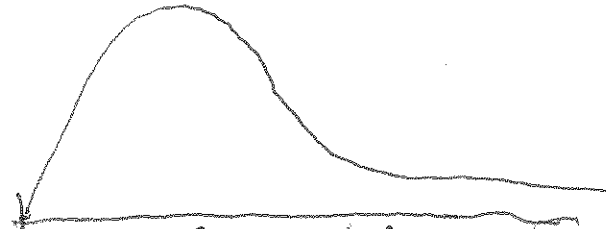
Do the data appear to fit some common/known distribution:



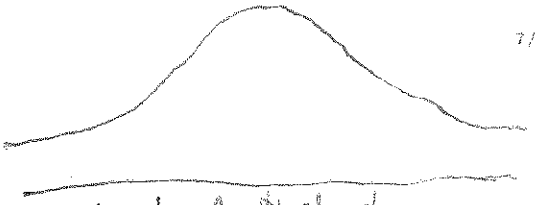
Normal / Gaussian







0 Chi-Square / Gamma



t-distribution

Visualizing 1-D Continuous Distributions

→ divide range into bins

How do we visualize continuous distributions?

Histogram ≈ Bar Chart
- For continuous data
- → count obs in each bin