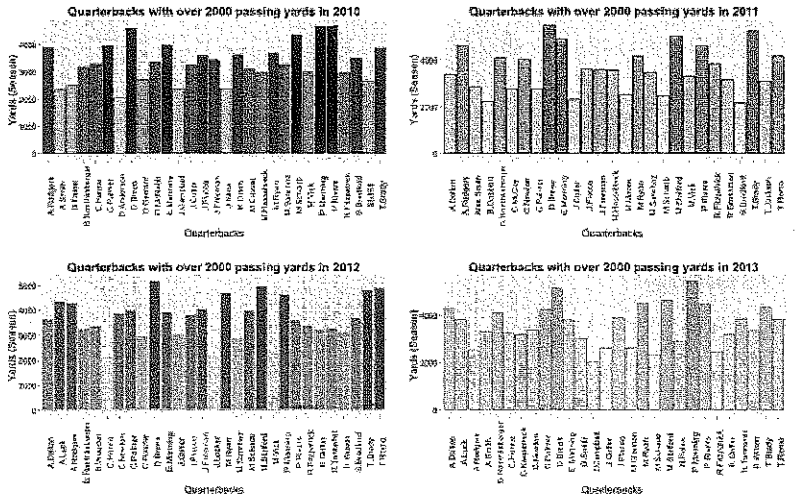
Today: 2-D Caterogical Data
Independence and Mosaic Plots

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

February 1, 2016

Quarterbacks with Over 2000 Yards, 2010 – 2013

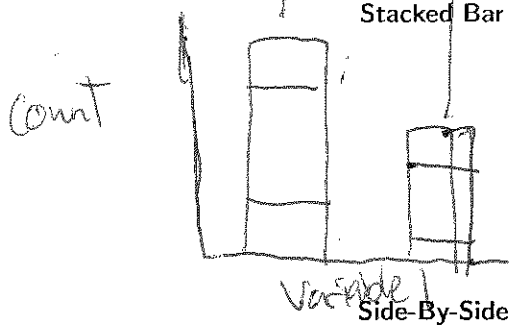


→ conditional distn of Var2 | Var1 = cat 1

→ cond. distn of Var2 | Var1 = cat 2

Last Time: 2-D Bar Charts

Stacked Bar Charts:



→ Marginal distn of Var 1

↳ easy to get info about marg.

→ difficult to get info about marginal Var 2

Side-By-Side Bar Charts:

→ cond. distn of V2 | V1 = 1

V2 | V1 = 2

→ hard to see marginal distn of Var 1 in side-by-side.

→ easy to see cond. distn of V2 | [redacted] V1

Contingency Tables

Two Categorical Variables:

↳ Give us
- Marginal distns (of one variable)
- Joint distns (of Two variables)
- cond. distns (Var 2 | Var 1, Var 1 | Var 2)

~~**Asympototics:**~~

Two Cat. variables:

Var 1 has K_1 categories

Var 2 has K_2 categories

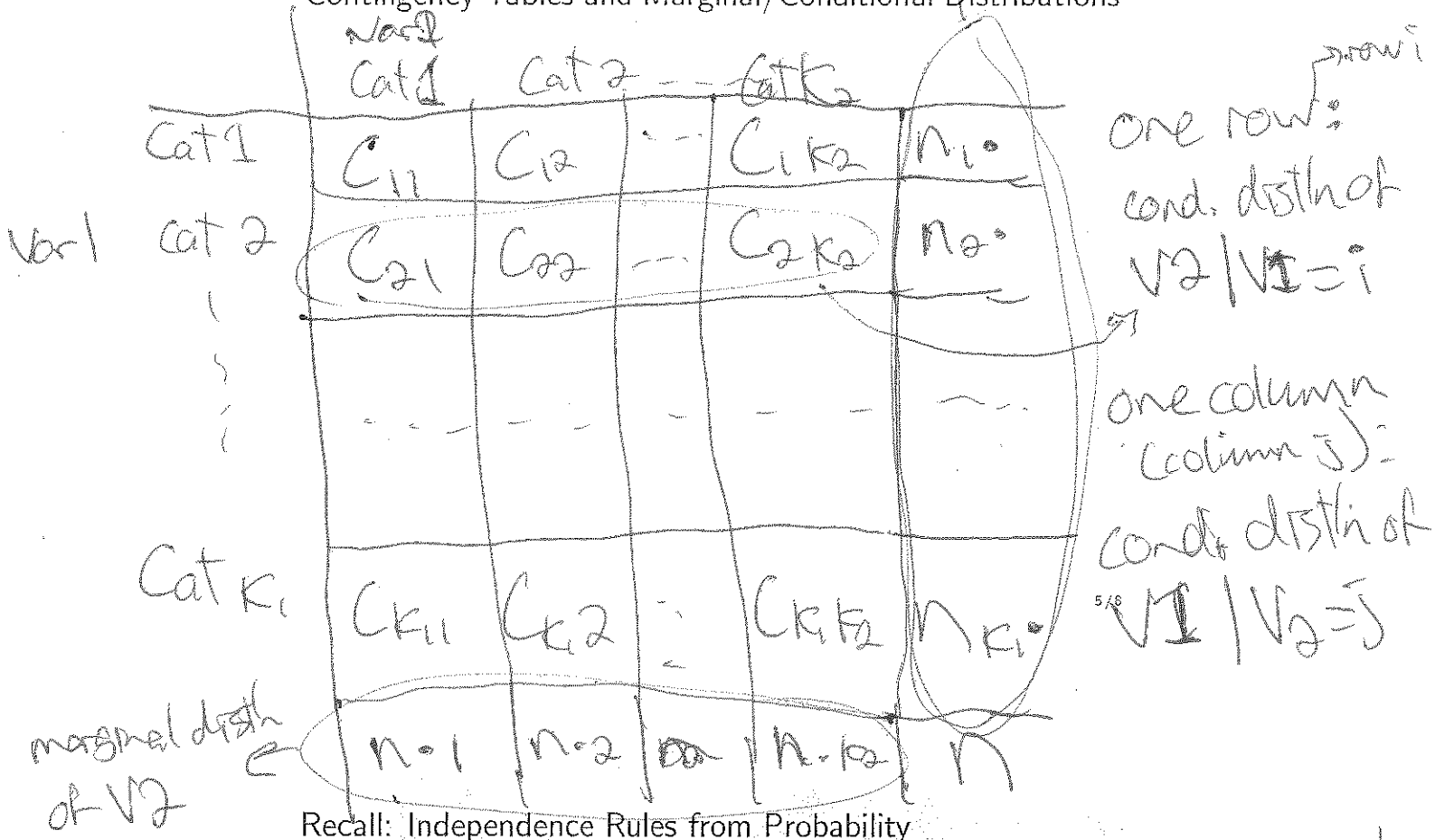
Interested in:

- observed counts in each cell
  - C_{ij} = # obs w/ var 1 = i, var 2 = j
- Expected counts in each cell
  - $\hat{C}_{ij}$ = [redacted] $n_{i\cdot} \times n_{\cdot j} / n$

whole thing = joint distn

marginal distn of Var 1

Contingency Tables and Marginal/Conditional Distributions



Recall: Independence Rules from Probability

$P(A|B) = P(A) \rightarrow$ conditional of $V_1|V_2$ = marginal of V_1

$P(B|A) = P(B) \rightarrow$ ——— $V_2|V_1$ = ——— V_2

$P(A \cap B) = P(A)P(B) \rightarrow$ joint = (marginal of V_1) × (marginal of V_2)

Can input contingency tables into chi-square tests for independence

E.g. `chisq.test(table(var1, var2))` → contingency table in R

More on this in Lab 04 → labs

Pearson Residuals → deviations in what we observe vs. what we expected.

Pearson Residuals: Scaled difference between observed/expected

$$r_{ij} = \frac{C_{ij} - \hat{C}_{ij}}{\sqrt{\hat{C}_{ij}}} = \frac{\text{Obs.} - \text{Exp.}}{\sqrt{\text{Exp}}}.$$

$r_{ij} > 0$: "too many" observed

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

Asymptotics

r_{ij}'s are asymptotically Normal!

$|r_{ij}| > 2 \Rightarrow$ signif. at $\alpha = 0.05$ level

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

Mosaic Plots → visualizes contingency tables

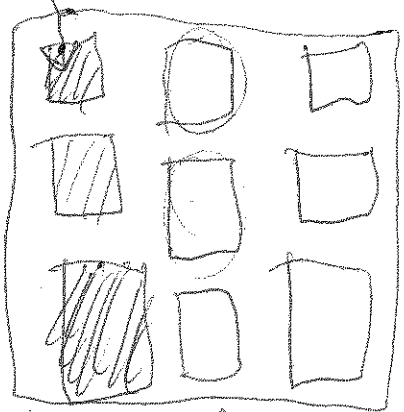
Mosaic Plots: Area plot for two categorical variables

each cell in contingency table gets a box in the plot

area ∝ % of obs. in the corresponding cell of the contingency table

width ∝ % in Var1 cat. i → marginal dist'n of Var1

height ∝ % in Var2 cat j | Var1 = cat i → conditional dist'ns



Can color the boxes by their differences from what was expected:

Wednesday: Mosaic Plot Demo