ST 518: Data Analytics II

Contingency Tables

Contingency Tables

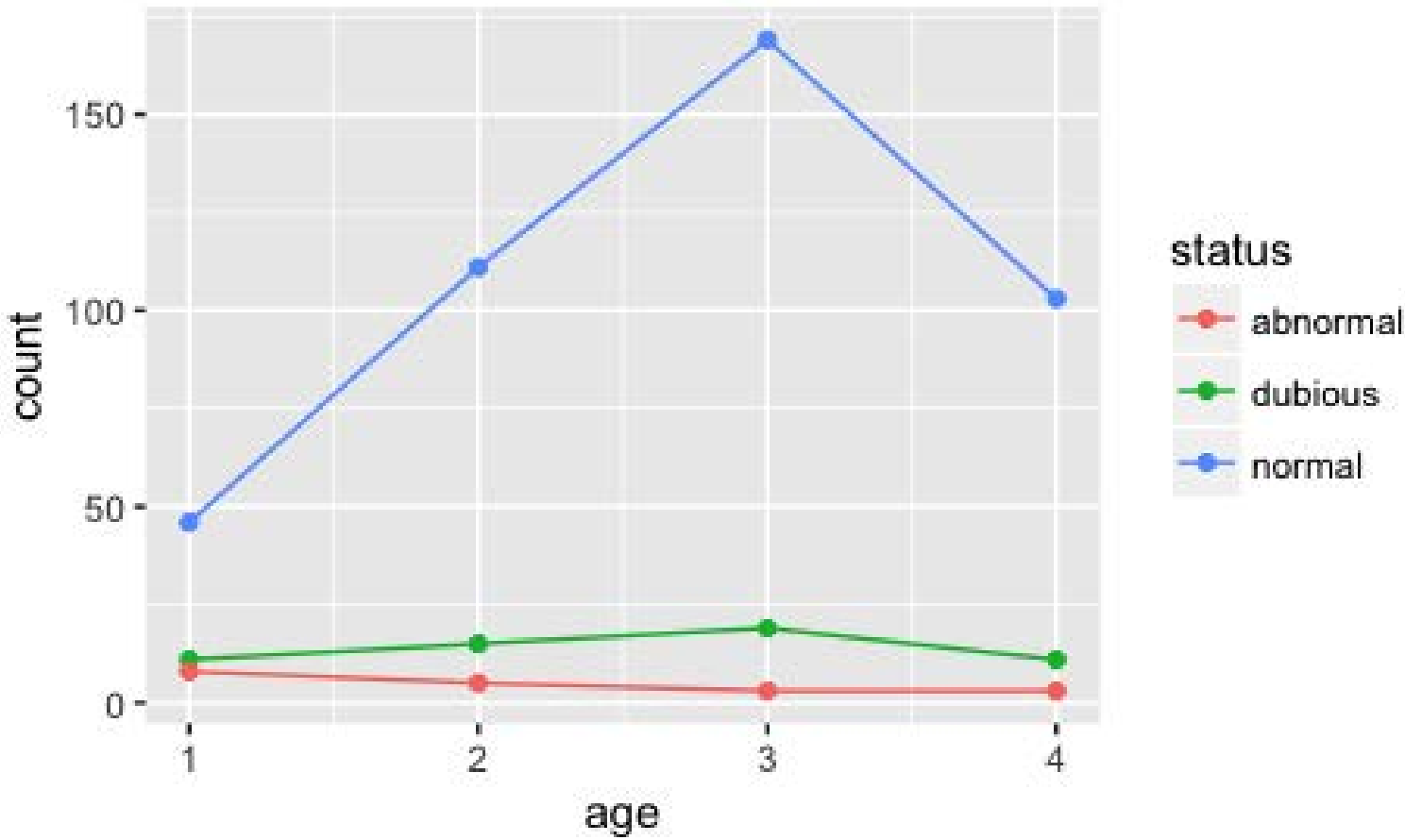
- Aggregation
- Simpson's Paradox

Displaying Categorical Data

Categorical data are often quite usefully displayed in tabular form. The data in Example 12.1 of *Statistical Methods* appear in a 3×4 contingency table—505 infants were cross classified by their Prechtl status (3 categories) and their gestational age at birth (4 categories):

	Gestational Age (weeks)			
Prechtl Status	≤ 31	32–33	34–36	≥ 37
Normal	46	111	169	103
Dubious	11	15	19	11
Abnormal	8	5	4	3

Use Plots Carefully



Graduate Admissions

In a famous example, there appeared to be striking evidence of a gender bias in admissions to graduate programs at the University of California at Berkeley (Bickel et al., 1975):

Gender	Admitted	Denied
Male	3738	4704
Female	1494	2827

These data show that 44% of males were admitted, whereas only about 35% of females were admitted, suggesting bias.

The Perils of Aggregation

These **aggregated** data don't tell the whole story, however. It turns out that if we split these data into six separate 2×2 tables, where the splitting depends upon department, we get a different story:

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Simpson's Paradox

The apparent paradox in the previous example—the aggregated data show a male gender bias, and some of the non-aggregated data show a female gender bias—is called Simpson's Paradox.

- The big issue to remember is to be careful about aggregating data—you can fail to see important relationships,
- When looking at data in contingency tables, try to think about what other factors might be missing from such a summary—in the graduate admissions example, it was important to think about the departments to which applicants applied. Are there other factors that should have been considered?