Exploring the cancer data

Hadley Wickham & Heike Hofmann

Cancer data

- Cancer incidence and mortality (+ population), broken down by:
  - 52 states (!!)
  - 2 sexes (male, female)
  - 3 races (black, hispanic, white)
  - 26 sites
  - 6 years (1999–2004)

Goals

- Practice data handling skills
- Practice asking interesting questions
- Investigate spatial and temporal patterns:
  - Time series plots
  - Chloropleth (map) plots
- Learn how to aggregate data with reshape

Example

```
library(reshape)

cancerm <- melt(cancer, id = 1:5)

cast(cancerm, race ~ variable, sum)

cast(cancerm, sex ~ variable, sum)

cast(cancerm, state ~ variable, sum)
```

Our first function

```
rates <- function(df) {
  transform(df,
    irate = incidence / population * 100000,
    mrate = mortality / population * 100000
  )
}
site_rates <- rates(cast(cancerm, site ~
variable, sum))
```

Cancer data

```
site_rates <- rates(
  cast(cancerm, site ~ variable, sum)
)

qplot(irate, site,
data=site_rates, xlim=c(0, NA))

qplot(irate, reorder(site, irate),
data=site_rates, xlim=c(0, NA))
```

Your turn

- Investigate the distribution of rates by state, race, sex, and year
- Are overall rates of cancer increasing or decreasing?
- What state has the highest overall cancer rate?

Your turn

- Investigate the distribution of rates by site and sex (hint: site + sex ~ variable). What cancers are particularly different between the sexes? What about between different races?
- Break down rates by state and time. Plot a time series and look for interesting trends.

Choropleth maps

- How can we show the spatial distribution of cancer rates?
- What exactly is a map?

```
states <- read.csv("states.csv")

qplot(x, y, data=states, geom="path",
group=state)
qplot(x, y, data=states, geom="polygon",
group=state)

map_rates <- merge(states, state_rates,
by="state")

qplot(x, y, data=map_rates, group=state,
fill=irate, geom="polygon")
qplot(x, y, data=map_rates, group=state,
fill=mrate / irate, geom="polygon")
```

Your turn

- Load states.csv into R
- Summarise the cancer data at the state level. Combine with the states data and plot.
- Can you find a cancer with a clear geographic trend? (Hint: Use `cast` to produce a summary by state and site, and `subset` to pull out a single site)
- Extra: look at http://had.co.nz/ggplot2/scale_gradient.html and experiment with different colour schemes