4/7/09 Math 152

- empirical cum. dist function.
  - explanation.
  - Kolmogorov Smirnow test.

R: Bees wax example

- Survival functions / hazard rate.
  - explanation.
  - Guinea Pig data.

Q-Q plots.
additive and multiplicative treatment effects.
group III and V from
Bjerkdahl Guinea Pig data.

Histograms., Density Curves, Stem and Leaf plots.

Measures of location.
- Arith mean
- median

$$\text{Var}\left(\log\left(1 - F_n(t)\right)\right) \approx \frac{1}{100} \frac{1 - e^{-t}}{e^{-t}}$$

$$= \frac{1}{100}\left(e^t - 1\right).$$

$$\text{sd} \approx \frac{1}{10}\sqrt{e^t - 1}.$$

e.c.d.f

$X_1, \ldots, X_n$ a batch of numbers.

$$F_n(x) = \frac{1}{n}\left(\#\{X_i \le x\}\right).$$

$$= \text{proportion of observations} \le x.$$

If $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ is the ordered batch of numbers.

$F_n(x) = 0$ for $x < X_{(1)}$

$= 1/n$ for $X_{(1)} \le x < X_{(2)}$

$\vdots$

$= k/n$ for $X_{(k)} \le x < X_{(k+1)}$

etc.

F_n is piecewise constant, with a jump of $\frac{1}{n}$ at each $X_{(k)}$.

(if the $X_{(k)}$ are not distinct the jump is a multiple of $1/n$).

We have

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$$

where $I(X_i \le t) = \begin{cases} 1 & X_i \le t \\ 0 & X_i > t. \end{cases}$

So $\mathrm{Var}\left(F_n(t)\right) = \frac{1}{n^2} \cdot n \cdot F(t)(1-F(t))$

$$= \frac{F(t)(1-F(t))}{n}.$$

$$\log(1-F_n(t)) \approx \log(1-F(t))$$

$$\approx -\frac{1}{1-F(t)}(F_n(t)-F(t)) + \cdots$$

$$\mathrm{Var}(\log(1-F_n(t))) \approx \frac{\mathrm{Var}(F_n(t))}{(1-F(t))^2}$$

$$= \frac{1}{n}\frac{F(t)(1-F(t))}{(1-F(t))^2}$$

$$= \frac{1}{n}\left(\frac{F(t)}{1-F(t)}\right).$$

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t) \qquad \text{where } I(X_i \le t) = \begin{cases} 1 & \text{if } X_i \le t \\ 0 & \text{if } X_i > t \end{cases}$$

$$\mathrm{Var}(F_n(t)) = \frac{1}{n^2} \cdot n\, F(t)(1-F(t)) = \frac{F(t)(1-F(t))}{n}.$$

$$D_n = \max_{-\infty < t < \infty} |F_n(t) - F(t)|, \quad \text{dist of } D_n \text{ is independent of } F.$$

and this provides a test for

H_0: $X_1, \ldots, X_n$ are i.i.d $\sim F$.

Hazard function

Survival function

$$S(t) = P(T > t) = 1 - F_T(t)$$

$$S_n(t) = 1 - F_n(t)$$

$$= \text{proportion of data greater than } t.$$

hazard function

instantaneous death rate for individuals who have survived up to a given time

$$P(t \le T \le t+\delta \mid T \ge t) = \frac{P(t \le T \le t+\delta)}{P(T \ge t)}$$

$$\cong \frac{\delta f(t)}{1 - F(t)}$$

The hazard function is defined as

$$h(t) = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log(1 - F(t))$$

$$= -\frac{d}{dt} \log S(t)$$

Since $\operatorname{Var}\left[\log(1 - F_n(t))\right] \cong \frac{1}{n} \frac{F(t)}{1 - F(t)}$

values of t s.t. $F(t)$ is near 1 must be discarded

→ See Calculation on next page.

$$\log\left(1 - F_n(t)\right) \approx \log\left(1 - F(t)\right)$$

$$- \frac{1}{1 - F(t)} \left(F_n(t) - F(t)\right) + \cdots$$

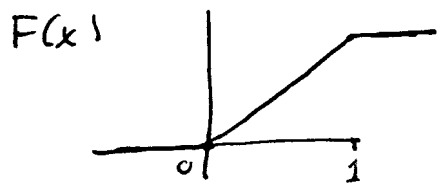
$$\mathrm{Var}\left(\log\left(1 - F_n(t)\right)\right) \approx \frac{\mathrm{Var}\left(F_n(t)\right)}{\left(1 - F(t)\right)^2}$$

$$= \frac{1}{n} \frac{F(t)\left(1 - F(t)\right)}{\left(1 - F(t)\right)^2}$$

$$= \frac{1}{n} \left(\frac{F(t)}{1 - F(t)}\right).$$

See problems 2 and 8.

2. $X_1, \ldots, X_n$ iid $U[0,1]$.



$$\sqrt{E\left((F_n(x) - x)^2\right)} = \sqrt{\frac{F_n(x)\left(1 - F_n(x)\right)}{n}}$$

8. $$\mathrm{Var}\left(\log\left(1 - F_n(t)\right)\right) \approx \frac{1}{100} \frac{1 - e^{-t}}{e^{-t}}$$

$$= \frac{1}{100}\left(e^t - 1\right)$$

so s.d. $\approx \frac{1}{10}\sqrt{e^t - 1}$

A surprising result (which we won't prove) is that the distribution of

$$D_n = \max_{-\infty < t < \infty} |F_n(t) - F(t)|$$

is independent of F.

This provides a test for

H_0: $X_1, \ldots X_n$ are i.i.d with c.d.f F.

. . .

R: Beeswax example.

e.g. Guinea Pig data.

as dosage increases
instantaneous mortality rates increase more ~~slowly~~ quickly
and reach higher levels.

increased mortality rate sets in earlier for higher dosage groups and seems greater.

notice the large dispersion as t increases again:

$$\text{Var}\{\log(1-F_n(t))\} \approx \frac{\text{Var}[1-F_n(t)]}{(1-F(t))^2}$$

$$= \frac{1}{n}\frac{(F(t))}{(1-F(t))}$$

so the variance becomes large as $F(t) \to 1$.

Q-Q plots

When making probability plots we plotted sample quantiles
(assigning the value $X_{(k)}$ to the $\frac{k}{n+1}$ quantile when there are n data points)
vs. the quantiles of a theoretical distribution
as a way of investigating how well the theoretical dist. fits the data.

To compare two batches of numbers, (which may come from the same or different distributions)
we will plot
the empirical quantiles of one batch
vs. the other

$X_{(1)} \ldots X_{(n)}$ vs. $Y_{(1)}, \ldots Y_{(n)}$.

plot $(X_{(i)}, Y_{(i)})$.

(if the batches are of different sizes, there
is an interpolation procedure).

e.g. (additive treatment effect).

(cdf's) F models a control group's data (x)

G models a treatment group's data. (y)

If the treatment ~~increases~~ changes the
expected response of each individual by
the same fixed amount h

and $F(x_p) = p = G(y_p)$

then we should expect $y_p = x_p + h$.

and the Q-Q plot will be a line
with slope 1 and intercept h.

We have $G(y_p) = F(y_p - h)$ so an additive
treatment effect corresponds to a shift of the cdf

multiplicative treatment effect.

response is multiplied by a constant c.

$$y_p = c x_p$$

and the Q-Q plot will be a line with slope c and intercept 0.

We have $G(y_p) = p = F(x_p) = F\left(\frac{y_p}{c}\right)$

so the treatment effect corresponds to a rescaling of the c.d.f.

e.g. Q-Q plot of III vs. V from the Guinea Pig data.