# Biostatistics 209, Lab #1

## 1. Background

**Data Description**: This is data on 90 males diagnosed with cancer of larynx during 1970-1978 at a Dutch hospital. Times are the interval between first treatment and death or end of study (01/01/93). Also recorded are the patient's age at diagnosis and stage of patient's cancer. The four stages are based on the *T.N.M.* (primary tumor *(T),* nodal involvement *(N)* and distant metastasis *(M)* grading) classification. Stage I: $T_1\ N_0\ M_0$, Stage II: $T_2\ N_0\ M_0$, Stage III: $T_3\ N_0\ M_0$ and $T_x\ N_1\ M_0$ (where $x$ = 1, 2, 3) and Stage IV: all other combinations.

| | | | |
|---|---|---|---|
| **Event Time**: | Time to death (in years) | `futime` | |
| **Event Type**: | Mortality Status | `died` | (0: censored, 1: death) |
| **Predictors**: | Disease Stage | `stage` | (I-IV) |
| | Age at Diagnosis | `age` | |

## 2. Data

- Download `larynx_M.dta` from the website
- Read in the data
  `use larynx_M.dta`
  *4 variables: stage, futime, age, died*
- Declare the data to be survival using `stset`
  `stset futime, failure(died)`
  *`Stata` will respond by calculating the # of failures, max, min and total follow-up time, etc.*
- What happens if you forget to specify the `failure` option?

In `Stata` survival data is declared by using `stset`. The time variable is given first; here it is `futime`. The event type variable is given with the failure option; here it is `died`. Typing the `stset` command given above says that values of `futime` where `died` is a non-zero number are actual death times. Other values of `futime` are taken by `Stata` to be censoring times.

## 3. Exploring Stage Effects Using Kaplan-Meier Curves

- Using the `sts graph` command to make Kaplan-Meier curves for the stages

  ```
  sts graph, by(stage)
  ```

- Save this graph to your disk so you can view it later. You can do this by right clicking on the graph.
- Try out options such as `failure`, `censored(single/number)`, `risktable`, or `ci`

  ```
  sts graph, by(stage) failure
  ```

  *This gives probability of death, rather than survival*

```
sts graph, by(stage) censored(single) risktable
```

*This adds a mark at each censoring time and lists number at risk at some time points*

```
sts g, by(stage) ci sep yline(.5) yline(.75, lc(green)) plotop(lc(blue)) ciop(lc(midblue) fc(none)) xtick(0(1)10)
```

*This shows 95% confidence intervals and can help to see median survival times and their CIs*

- **Question 3.1**: Based on the Kaplan-Meier's what is your impression of the influence of the stages on death? Does it appear that the effect of 1 unit change in stage is the same across the range of values?

- Use `sts list` to get the Kaplan-Meier values at years 1, 2 and 5

```
sts list, by(stage) at(1 2 5)
```

- Use `stci` to obtain median survival or other percentiles

```
stci, by(stage)        or      stci, by(stage) p(25)
```

## 4. The Cox Model

Cox models are fit using the `stcox` command followed by a list of predictors. The outcome is not specified but is implied by the `stset` command, which was given in Section 1. To fit a Cox model with a categorical predictor, we use the `i.` prefix. The group with lowest values of the predictor is selected as the reference by default.

- Fit a Cox model for stage

```
stcox  i.stage
```

**Question 4.1.:** Is stage a statistically significant predictor? Which stage is at highest risk of death? Which are second and third?

**Question 4.2.:** Does it appear that the effect of 1 unit change in stage is the same across the range of values?

**Question 4.3.:** Do your answers above agree with the Kaplan-Meier graphs?

  - You can calculate the hazard ratio between any two groups using the `lincom` command. To get the HR of stage III with respect to stage II, type

```
lincom 3.stage - 2.stage, hr
```

**Question 4.4.:** Obtain the hazard ratio of stage IV compared with Stage III.

**Question 4.5.:** Implement a trend test for stage:

```
contrast p.stage
```

Is there evidence of a linear trend?

## 5. Changing the Reference

In any multiple categorical variable, we need to choose a single category as the reference to which the other categories are compared. The "`i.`" syntax by default takes the lowest numerical category (here that is 1=stage 1) as the reference. It is possible to change the reference category. Try this and see how the results compare. The syntax to change the reference for `stage` to be stage IV (stage=4) is

```
stcox b(4).stage
```

- Alternatively you can declare the base level (instead of being model specific) as follows
  ```
  fvset base 4 stage
  stcox i.stage
  ```

  Note that the `fvset base` information will be stored with the dataset. To clear it,
  ```
  fvset clear stage
  ```

**Question 5.1.** Fit a new model for stage using 4 as the baseline group. How does it compare to the previous model with 1 as the baseline? Is stage a stronger predictor?

**Question 5.2.** Obtain the hazard ratio of stage III compared with Stage II. How does it compare to the one from Question 4.3?

**Question 5.3.** Obtain the hazard ratio of stage IV compared with Stage III. How does it compare to the one from Question 4.4?

**Question 5.4.** Overall, what is similar and different between the model fits with two different reference groups?

## 6. Continuous Predictors

**Question 6.1.:** What is the effect of age on survival after adjusting for stage?
```
stcox age i.stage
```

**Question 6.2.:** Obtain the hazard ratio of a 10 years increase in age. Has the significance level of the age effect become stronger?
```
lincom 10*age, hr
```

**Question 6.3.:** Obtain the hazard ratio of a 10 years decrease in age.
```
lincom -10*age, hr
```

**Question 6.4.:** Verify that the hazard ratio and limits of its confidence interval in Question 6.3 is the reciprocal (one divided by the value) of the corresponding values in Question 6.2.