BMI 713 / GEN 212

Lecture 10, Part I: Survival Analysis

- Kaplan-Meier curve
- Log-rank Test
- Cox PH Regression

November 18, 2010

Logistic Regression

- In linear regression, the response variable *Y* is continuous

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$
$$e \sim N(0, \sigma^2)$$

- We are interested in identifying explanatory variables that help us to predict the mean value of the response by explaining the observed variation in the outcomes
- However, we often have a response variable *Y* that is **dichotomous** rather than continuous

Logistic Regression

- We cannot simply use linear regression with 0/1 as outcome
- We define the 'logit' transformation for the probability of the outcome variable being a "1" (vs"0")

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

- logit(p) can take any value and we can perform regression
- How do we interpret coefficients?
- If *X* is a binary variable, β is the log of the odds ratio $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

Introduction to Survival Analysis

- **"Time-to-event" data:** In some studies, the response variable of interest is the length of time between an initial observation and the occurrence of a subsequent event
- Despite the name, "survival" analysis isn't only for analyzing time until death. It deals with any situation where the quantity of interest is amount of time until study subject experiences some relevant endpoint.
- This subsequent event is often called a "failure"; the terms "failure" and "event" are used interchangeably for the endpoint of interest.

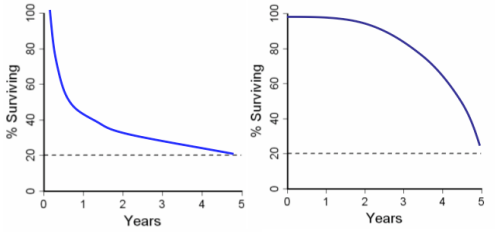
Survival Analysis

- Examples:
  - Time from birth until death
  - Time from start of treatment until remission of disease
  - Injection of a lentivirus with a growth factor into mice till the development of tumor
- The time from the initial event until failure is called the survival time
- Time is a continuous measurement that cannot assume negative values — it is rarely normally distributed
- We will study estimation, one-sample/two-sample inference, and regression in this context

How to Measure Survival

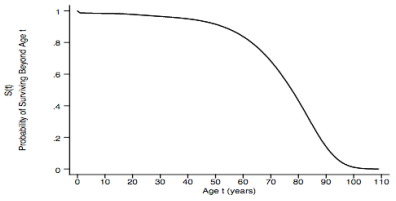
- **Idea:** Report the mean survival time?
- **Problem:** Not robust to outliers
- **Idea:** 5-year survival rate? "long" vs "short" survival
- **Problem:** how to choose the cutoff time?

- These two have the same 5-yr survival rates:



Survival Function

- A distribution of survival times can be characterized by a survival function ***S(t)***
- *S(t)* is the probability that an individual survives beyond time *t*, or the proportion of subjects who have not yet failed
- If *T* is a continuous random variable representing survival time, then *S(t) = P(T>t)*
- The graph of the survival function *S(t)* versus time *t* is called a survival curve



The Kaplan-Meier Survival Curve

- The Kaplan-Meier method, also known as the product-limit method, can be used to estimate a survival curve
- It is a **nonparametric** technique that does not make any assumptions about the underlying distribution of survival times
- **Example:** A total of 12 mice with a particular genotype were exposed to radiation and were followed until death

The Kaplan-Meier Curve

- Based on this sample, what can we infer about the survival?
- We begin by ordering the survival times associated with each of the 12 mice
- Again, 'survival' is a general term
- Another example: the mice with tumor were treated with a small molecule and were followed to remission

Mouse	Survival (weeks)
1	2
2	3
3	6
4	6
5	7
6	10
7	15
8	15
9	16
10	27
11	30
12	32

Survival Function

- Note that no one fails at 0 week or at 1 week following exposure, one mouse fails at 2 weeks, one at 3 weeks, none at 4 weeks, and so on
- N_t is defined as the number of mice who have not yet failed at time t:
  - $N_0 = 12$, $N_1 = 12$, $N_2 = 11$, …
- If everyone in the sample fails, then the survival function is estimated by $S(t) = N_t / N_0$

Survival Function

- Therefore,

  S(0) = 12/12 = 1.000

  S(2) = 11/12 = 0.917

  S(3) = 10/12 = 0.833

  S(6) = 8/12 = 0.667

  ….

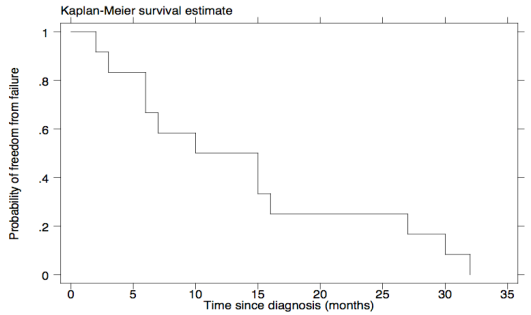
  S(32) = 0/12 = 0.000

- The Kaplan-Meier estimate of the survival curve is:

Time	N_t	*S(t)*
0	12	1.000
2	11	0.917
3	10	0.833
6	8	0.667
7	7	0.583
10	6	0.500
15	4	0.333
16	3	0.250
27	2	0.167
30	1	0.083
32	0	0.000

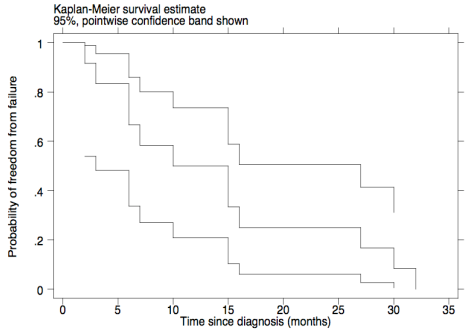
Survival Function

- *S(t)* is an estimate of the true population survival function calculated using the information in a sample of observations



Confidence Bands

- To quantify the amount of sampling variability involved, we can calculate the standard error of *S(t)* and use it to construct confidence bands



Survival Function

- Another way: $S(t_i)$ = P(alive at t_i | alive at t_{i-1}) x $S(t_{i-1})$
- Probability you survive until time t_i = probability you survive until t_{i-1} and then survive until t_i given you made it to t_{i-1}
- P(alive at t_i | alive at t_{i-1}) = (# alive at t_i)/(# alive at t_{i-1})

S(0) = 12/12 = 1.000	S(0) = 12/12 = 1.000
S(2) = 11/12 = 0.917	S(2) = 11/12 = 0.917
S(3) = 10/12 = 0.833	S(3) = 10/11 * S(2) = 10/11 * 11/12 = 0.833
S(6) = 8/12 = 0.667	S(6) = 8/10 * S(3) = 8/10 * 10/11 * 11/12 = .0667
….	….
S(32) = 0/12 = 0.000	S(32) = 0/12 = 0.000

Missing Data: Censoring

- Time-to-event data often have data missing in a particular way: **individuals may be lost to follow-up or may drop out of the study before they experience the event of interest**
- This incomplete observation of time to failure is known as **censoring**
- Censored data provide partial information: you do not know how long a patient lived, but you know that she/he lived at least as long as the time before being lost to follow-up.
- Why would a person be lost to follow-up? The person could have, e.g., moved to another city, withdrawn from the study, or died of a different cause.
- The data might be analyzed before the event of interest has occurred in all subjects

Censoring

- We would like to be able to take advantage of the partial information contained in censored observations
- To do inference in the setting of missing data, we must be willing to make a big assumption that **censoring is non-informative**
- In other words, assume that being lost to follow-up is unrelated to prognosis
- If this assumption can't be made, inference becomes more complicated if not impossible
- If the reason a person is lost to follow-up is related to prognosis, then our data is **biased**

Informative Censoring

- **Example:** Researchers administer a new chemotherapy drug to 10 cancer patients to estimate survival time while on the drug
- 5 patients can't tolerate the side effects and drop out of the study
- If non-informative censoring were assumed, the drug would probably appear falsely impressive.
- Those who dropped out were probably more ill; hence shorter survival times were disproportionately removed from the sample.

Kaplan-Meier Estimator

- The Kaplan-Meier method can be modified to account for the partial information about survival times that is available from censored observations
- Example: Suppose that, in the sample of 12 mice, mice 2 and 6 have not yet died
- Instead, they are alive after 3 and 10 months of follow-up, respectively, but are lost to follow-up

Mice	Survival (weeks)
1	2
2	3+
3	6
4	6
5	7
6	10+
7	15
8	15
9	16
10	27
11	30
12	32

Kaplan-Meier Estimator

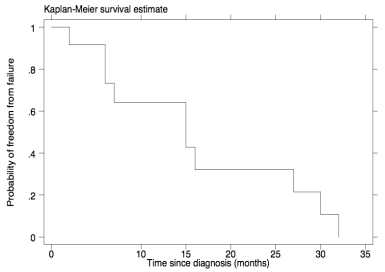
- $S(t_i) = P(\text{alive at } t_i \mid \text{alive at } t_{i-1}) \times S(t_{i-1})$
- If there were no censoring, $P(\text{alive at } t_i \mid \text{alive at } t_{i-1}) = (\#\text{ alive at } t_i)/(\#\text{ alive at } t_{i-1})$
- However, a patient who is alive but censored at time t_{i-1} never really had a chance to make it to t_i. That patient was not eligible to die during the interval from t_{i-1} to t_i and therefore should not be counted for computing survival rate in this interval.

S(0) = 12/12 = 1.000
S(2) = 11/12 = 0.917
S(3) = 10/11 * 11/12 = 0.833
S(6) = 8/10 * S(3) = 8/10 * 10/11 * 11/12 = .667

S(0) = 12/12 = 1.000
S(2) = 11/12 = 0.917
S(3) = S(2)
S(6) = 8/10 * S(2) = 8/10 * 11/12 = .733

Kaplan-Meier Estimator

- In this case, $S(t)$ does not change from its previous value if the observation at time t is censored $S(3)=S(2)=0.917$
- However, the observation is not used to calculate the probability of failure at any subsequent time–it is removed from the denominator



KM estimator in R

- Data: 2, 3+, 6, 6, 7, 10+, 15, 15, 16, 27, 30, 32

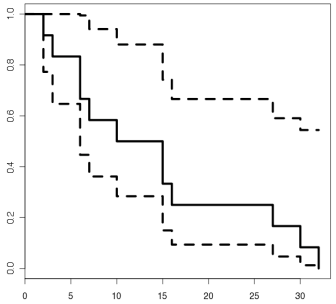
```
> library(survival)
> surv = c(2,3,6,6,7,10,15,15,16,27,30,32)
> status = c(1,0,1,1,1,0,1,1,1,1,1,1)
> Surv(surv,status)
 [1]  2   3+  6   6   7  10+ 15  15  16  27  30  32
> surv.all = survfit(Surv(surv,status))
> summary(surv.all)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     12       1    0.917  0.0798        0.7729        1.000
    6     10       2    0.733  0.1324        0.5148        1.000
    7      8       1    0.642  0.1441        0.4132        0.996
   15      6       2    0.428  0.1565        0.2089        0.876
   16      4       1    0.321  0.1495        0.1287        0.800
   27      3       1    0.214  0.1325        0.0635        0.720
   30      2       1    0.107  0.1005        0.0169        0.675
   32      1       1    0.000      NA            NA           NA
```

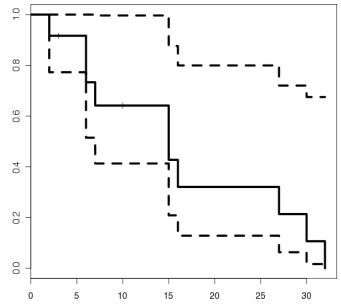
KM estimator in R

```
> plot(surv.all)
```

Without censoring



With censoring



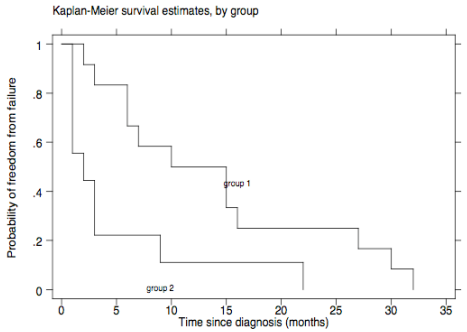
Log-Rank Test

- We often want to compare the distributions of survival times in two (or more) different populations to determine whether survival differs between the groups

Group 1		Group 2	
Patient	Survival (months)	Patient	Survival (months)
1	2	1	1
2	3	2	1
3	6	3	1
4	6	4	1
5	7	5	2
6	10	6	3
7	15	7	3
8	15	8	9
9	16	9	22
10	27		
11	30		
12	32		

Comparison of Two Groups

- It appears that mice in group 1 survive longer than those in group 2



Log-Rank Test

- We can test the null hypothesis that the two (or more) distributions of survival times are identical using a technique called the log-rank test
- ***$H_0 : S_1(t) = S_2(t)$***
- This test compares the observed number of failures at time *t* to the expected number of failures (assuming that the two curves are identical), and then accumulates the information over all times
- Under the null hypothesis, the test statistic has a chi-square distribution with 1 df
- Since p = 0.013, we reject H_0 and conclude that the two distributions of survival times are not identical

Log-Rank Test

- How does the log-rank test work?
- A direct application of the **Mantel-Haenszel test,** which combines data from a series of 2x2 tables
- Example: exposure/no exposure vs death/no death divided by another variable: sex, genotype, etc.
- The study period is subdivided into *k* intervals. For each interval, a 2x2 table is created. The test statistic is calculated from the *k* tables just as in Mantel-Haenszel
- The only extra thing to worry about is to remove the censored cases in between intervals.

Cox PH Model

- We are often interested in the relationship between survival time and a continuous risk factor, or to evaluate the simultaneous effects of more than one risk factor
- Log-rank test is only for one dichotomous variable
- Multivariate analysis can be performed using the **Cox proportional hazards model**
- Multiple linear regression analysis cannot be used because survival time is rarely normally distributed, and because it cannot account for censored observations
- The Cox model is an example of a **semiparametric** model

Cox PH Model

- We need a new function called the **hazard function, *h(t)***
- This is the probability that you will die in the very instant after time *t*, given that you have survived until time *t*
- The proportional-hazards model assumes that the hazard rate for any individual can be modeled as a function of covariates $X_1, \ldots, X_k$ as follows:

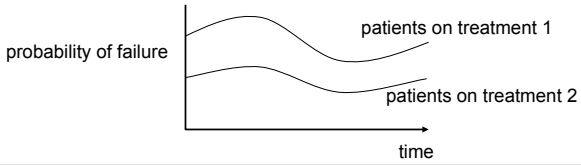
$$h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \cdots + \beta_k x_k$$

Cox PH Model

$$h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

- **$h_0(t)$** is called the "**baseline hazard rate**"
- We make no assumptions about its shape.
- This is why the model is called **semiparametric.** We don't completely specify the distribution of survival times; we only specify that changes in covariates will change the hazard rate proportionally to whatever it was.



Interpretation of the Coefficients

- Interpreting the parameters of the model is a bit difficult. The easiest case to understand is when a variable is dichotomous.
- Example: Suppose we are analyzing survival times using a Cox proportional hazards model with covariates X_1 = gender (1=F), X_2 = drug dosage. What is the ratio of hazards between a man and a woman on the same dose of the drug?
- $$\frac{h_{woman}(t)}{h_{man}(t)} = \frac{h_0(t)e^{\beta_1(1)+\beta_2 x_2}}{h_0(t)e^{\beta_1(0)+\beta_2 x_2}} = e^{\beta_1}$$
- β_1 is the logarithm of the **"hazard ratio"**, which can be thought of as the instantaneous relative risk of death per unit time of a woman vs. of a man, given that both have survived until time t and with all other covariates held constant

Summary

- **Survival analysis** to handle survival data which usually have censored data points and are non-normally distributed
- **Kaplan-Meier estimator** for estimation & one-sample inference
- **Log-Rank test** for Two-sample comparisons
- **Cox Proportional Hazards model** for regression modeling

Regression Models

- Summary:
  - binary (disease vs. normal) → Logistic regression (and many others!)
  - discrete
    - non-ordered (multiple subclasses) → Polytomous regr
    - ordered (number of recurrences) → Poisson regression
  - continuous (gene expression) → Linear regression
  - censored (patient survival time) → Cox model

Lecture 10, Part II: Multiple Testing

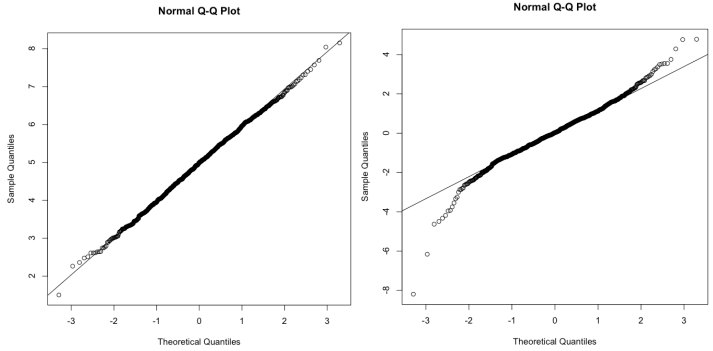
- Permutation Test
- Testing for Normality
- Family-Wise Error Rate
- False Discovery Rate

Permutation Tests

- Analyze the problem: think carefully about the null and alternative hypotheses
- Define a test statistic
- Calculate the test statistic for the original labeling of the observations
- Permute the labels and recalculate the test statistic
  - Exact Test
  - **Monte Carlo Test**
- Calculate p-value by comparing where the observed test statistic value lies in the distribution of test statistics under permutation

Testing for Normality: qqplot

```
x = rnorm(1000); qqnorm(x); qqline(x)
x = rt(1000,df=5); qqnorm(x); qqline(x)
```



Kolmogorov-Smirnov Test

- A non-parametric test for comparing two distributions

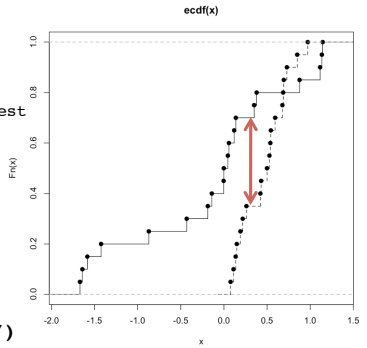
```
> x = rnorm(20)
> y = runif(20)
> ks.test(x,y)

   Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.6, p-value = 0.001116
alternative hypothesis: two-sided

> plot.ecdf(x,verticals=T)
> lines(ecdf(y),verticals=T, lt=2)

> ks.test(rnorm(100),"pnorm")
```



Errors in hypothesis testing

- Type I and Type II errors

	H_0 is true	H_1 is true
Reject H_0	Type I error	correct
Not reject H_0	correct	Type II error

- P(Type I error) = P(reject H_0 | H_0 is true) = α "false alarm"
- P(Type II error) = P(not reject H_0 | H_1 is true) = β "alarm failure"

- Power = P(reject H_0 | H_1 is true) = $1 - \beta$

Multiple Testing

- Large-scale experiments lead to lots of hypotheses
- A typical genome-wide experiment might result in performing 20000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect 1000 genes to be deemed "significant" by chance.

- If we perform *m* hypothesis tests, what is the probability of at least 1 false positive?

  P(Making an error) = α

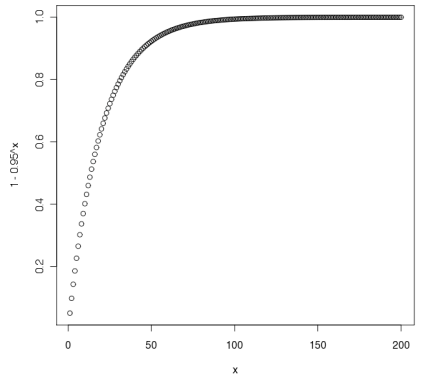
  P(Not making an error) = $1 - \alpha$

  P(Not making an error in *m* tests) = $(1 - \alpha)^m$

  P(Making at least 1 error in *m* tests) = $1 - (1 - \alpha)^m$

Multiple Testing

- Probability of at least one false positive: $1 - (1 - \alpha)^m$



Family-Wise Error Rate

- One option is to control this Family-Wise Error Rate (the probability of at least one type I error)
- We already saw the Bonferroni method: simply divide α by the number of tests
- If we want to have α=0.05 when we perform 10,000 tests, use a p-value of $0.05/10000 = 5 \times 10^{-6}$ as the threshold for significance
- We saw that this is too conservative; probability of Type II errors is too high

False Discovery Rate

- In many large-scale experiments, **we can tolerate some false positives**
- FWER is appropriate when you want to guard against ANY false positives
- Thus a popular alternative is control the false discovery rate (FDR)

FDR

	H_0 is true	H_1 is true	Total
Reject H_0	**V**	**S**	R
Not reject H_0	**U**	**T**	m - R
	m_0	$m-m_0$	m

- **FDR = V/R**
- FDR is designed to control the **proportion of false positives among the set of rejected hypotheses** rather than to control the Type I error rate

Benjamini-Hochberg FDR

- To control FDR at level q:
- Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$ for hypotheses $H_1, H_2, \ldots, H_m$.
- Find the test with the highest rank k for which the p-value is less or equal to $(k/m) * q$
- Reject all H_i for $i \leq k$
- On the right: q=0.05; m=10

Rank (k)	p-value	$(k/m) * q$	Reject
1	0.003	0.005	R
2	0.008	0.010	R
3	0.012	0.015	R
4	0.021	0.020	0
5	0.070	0.025	0
6	0.123	0.030	0
7	0.250	0.035	0
8	0.673	0.040	0
9	0.812	0.045	0
10	0.890	0.050	0

FDR vs pFDR

- **When the test statistics are independent, this procedure controls the FDR at the level *q*.**

Technical Details:

- Actually FDR $\leq q*m_0/m$. We can try to estimate m_0/m
- Also true under positive and negative correlations
- For highly correlated data, this may be conservative; use more powerful FDR procedure by resampling

- Benjamini-Hochberg: FDR = E[V/R | R>0] P(R>0)
- Storey & Tibshirani: pFDR = E[V/R | R>0]
- P(R>0) ~ 1 in nearly all cases and so the two are very similar

q-value

- q-value is the minimum FDR that can be attained when calling that feature significant (i.e., **expected proportion of false positives incurred when calling that feature significant**)
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)
- Example: In a microarray study for differential expression, if gene X has a q-value of 0.04 it means that 4% of genes that show p- values less than or equal to that of gene X are false positives
- These q-values are still **estimates**

q-values in R

```
> install.packages("qvalue")
> library(qvalue)
> qv = qvalue(my.p.values)
> names(qv)
[1] "call"    "pi0"     "qvalues" "pvalues" "lambda"
> qv$qvalues
 [1] 7.716203e-01 5.869116e-01 7.874598e-01
   6.701546e-018.442438e-01 8.536429e-01 1.930196e-01
   6.788870e-01
```

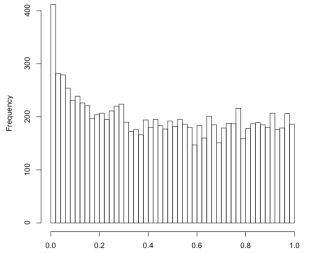
R Simulations

- Suppose I would like to compare the means of two groups. H_0: the mean is zero vs H_1: the mean is not zero
- Generate 9000 true hypotheses
- Generate 1000 false hypotheses
- For each hypothesis, draw 10 samples
  - for H0: rnorm(10,mean=0)
  - for H1: rnorm(10,mean=3)
- Compute p-values from t-statistic
- Using alpha = 0.05 → **FP: 436, FN: 0**
- Using alpha/10000 (Bonferroni) → **FP: 0, FN: 228**
- Using FDR = 0.05 → **FP: 54, FN: 0**

Multiple Testing Corrections

- So what is the procedure in practice?
- Should the significance my gene depend on that of other genes?

Plot the distribution of p-values



- What is the proper threshold for q-values?