A Short Course on Biostatistics

Part II: Inference

Babak Shahbaba, Ph.D.

Associate Professor, Department of Statistics
University of California, Irvine

July 31, 2017

Overview

- We now focus on formal statistical inference:
    - Estimation
        - Point estimation
        - Interval estimation
    - Hypothesis testing
        - One-sample t-test
        - Two-sample t-test
        - Nonparametric tests
        - Analysis of Variance (ANOVA)
        - Analysis of categorical variables
- At the end, we will briefly discuss regression models and survival analysis

Estimation

Parameter estimation

- We are interested in the **population mean** and **population variance**, denoted as μ and σ^2 respectively, of a random variable.
- These quantities are **unknown** in general.
- We refer to these unknown quantities as **parameters**.
- We discuss statistical methods for parameter **estimation**.
- Estimation refers to the process of guessing the unknown value of a parameter (e.g., population mean) using the observed data.

Convention

- We denote the unknown population mean and variance as μ and σ^2 respectively.
- We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population.
- We treat $X_1, X_2, \ldots, X_n$ themselves as n random variables because their values can change depending on which n individuals we sample.
- We assume the samples are **independent and identically distributed** (IID).
- We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample.

Point estimation vs. interval estimation

- Sometimes we only provide a single value as our estimate.
- This is called **point estimation**.
- We use $\hat{\mu}$ and $\hat{\sigma}^2$ to denote the point estimates for μ and σ^2.
- Point estimates do not reflect our **uncertainty**.
- To address this issue, we can present our estimates in terms of a range of possible values (as opposed to a single value).
- This is called **interval estimation**.

Estimating population mean

- Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:

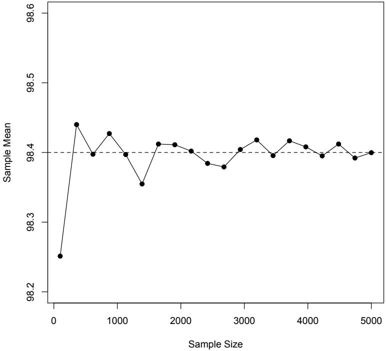
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- In this case, we say that $\bar{X}$ is an **estimator** for μ.
- The estimator itself is considered as a random variable since it value can change.
- We usually have only one sample of size n from the population $x_1, x_2, \ldots, x_n$.
- Therefore, we only have one value for $\bar{X}$, which we denote $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Law of Large Numbers (LLN)

- Suppose the true population mean for normal body temperature is 98.4F.



- Here, the estimate of the population mean is plotted for different sample sizes.

Estimating population variance

- Given n randomly sampled values $X_1, X_2, \ldots, X_n$ from the population and their corresponding sample mean $\bar{X}$, we estimate the population variance as follows:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

- The sample standard deviation S (i.e., square root of S^2) is our estimator of the population standard deviation σ.
- We regard the estimator S^2 as a random variable.
- In practice, we usually have one set of observed values, $x_1, x_2, \ldots, x_n$, and therefore, only one value for S^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Confidence intervals for the population mean

- It is common to express our **point estimate along with its standard error** to show how much the estimate could vary if different members of population were selected as our sample.
- Standard error is calculated as

$$\mathrm{SE} = s/\sqrt{n}$$

- Alternatively, we can use the point estimate and its standard deviation to express our estimate as a range (interval) of possible values for the unknown parameter.
- This is known as **confidence interval**.
- To find confidence intervals, we need to know the probability distribution of the corresponding estimator.
- Probability distributions for estimators are called **sampling distributions**.

Confidence intervals for the population mean

- Here, we are mainly interested in the **sampling distribution** of $\bar{X}$, for which we have (either exactly or asymptotically through the *central limit theorem*) we have the following results.
- With 95% *probability* $\bar{X}$ is in

$$\left[\mu - 2 \times \sigma/\sqrt{n},\ \mu + 2 \times \sigma/\sqrt{n}\right].$$

- But we are interested in μ: with 95% *probability* μ is in

$$\left[\bar{X} - 2 \times \sigma/\sqrt{n},\ \bar{X} + 2 \times \sigma/\sqrt{n}\right].$$

- However we have only one $\bar{x}$ from the observed data, so we refer to this as our 95% *confidence interval*

$$\left[\bar{x} - 2 \times \sigma/\sqrt{n},\ \bar{x} + 2 \times \sigma/\sqrt{n}\right].$$

Confidence intervals for the population mean

- Also, since σ is unknown, we need to estimate it using s. If the sample size is large enough, we can still use the following interval as our 95% confidence interval:

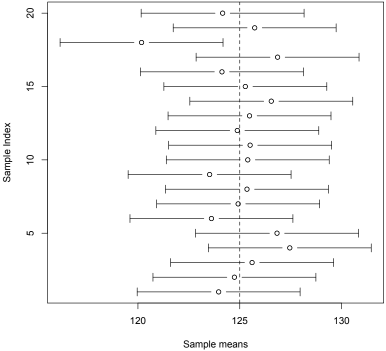
$$\left[\bar{x} - 2 \times s/\sqrt{n},\ \bar{x} + 2 \times s/\sqrt{n}\right].$$

- In general, the confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - t_{\mathrm{crit}} \times SE,\ \bar{x} + t_{\mathrm{crit}} \times SE\,\right]$$

where the factor t_{crit} is obtained from a t-distribution with $n - 1$ degrees of freedom.

Interpretation of confidence interval



Margin of error

- We can write the confidence interval as

$$\bar{x} \pm 2 \times SE$$

- The term $2 \times SE$ is called the **margin of error** for 0.95 confidence level.
- In general, for a given confidence level, margin of error is $t_{\mathrm{crit}} \times SE$.
- It is common to present interval estimates for a given confidence level as

$$\text{Point estimate} \pm \text{Margin of error.}$$

Hypothesis testing

Hypothesis

- In general, many scientific investigations start by expressing a hypothesis.
- For example, Mackowiak et al (1992) hypothesized that the average normal (i.e., for healthy people) body temperature is less than the widely accepted value of $98.6F$.
- If we denote the population mean of normal body temperature as μ, then we can express this hypothesis as $\mu < 98.6$.

Null and alternative hypotheses

- The null hypothesis usually reflects the “status quo” or “nothing of interest”.
- In contrast, we refer to our hypothesis (i.e., the hypothesis we are investigating through a scientific study) as the **alternative hypothesis** and denote it as H_A.
- For hypothesis testing, we focus on the null hypothesis since it tends to be simpler.

Null and alternative hypotheses

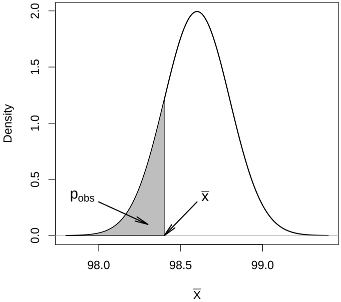
- With respect to our decision regarding the null hypothesis H_0, we might make two types of errors:
  - Type I error: we reject H_0 when it is true and should not be rejected.
  - Type II error: we fail to reject H_0 when it is false and should be rejected.
- We denote the probability of making type I error as α and the probability of making type II error as β.
- We refer to $1 - \beta$ (i.e., correctly rejecting the null hypothesis) as the **power** of the test.

Example

- Consider the body temperature example, where we want to examine the null hypothesis $H_0 : \mu = 98.6$ against the alternative hypothesis $H_A : \mu < 98.6$.
- Suppose that we have randomly selected a sample of 25 healthy people from the population and measured their body temperature.
- To decide whether we should reject the null hypothesis, we quantify the empirical support (provided by the observed data) against the null hypothesis using some statistics.
- We use statistics to evaluate our hypotheses.
- We refer to them as **test statistics**.
- For a statistic to be considered as a test statistic, its sampling distribution must be fully known (exactly or approximately) under the null hypothesis.
- We refer to the distribution of test statistics under the null hypothesis as the **null distribution**.

Observed significance level

- The **observed significance level** for a test is the probability of values as or more extreme than the observed value, based on the null distribution in the direction supporting the alternative hypothesis.
- This probability is also called the ***p*-value** and denoted p_{obs}.



Interpretation of *p*-value

- The *p*-value is the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true.
- When the *p*-value is small, say 0.01 for example, it is rare to find values as extreme as what we have observed (or more so).
- As the *p*-value increases, it indicates that there is a good chance to find more extreme values (for the test statistic) than what has been observed.
- Then, we would be more reluctant to reject the null hypothesis.
- A common **mistake** is to regard the *p*-value as the **probability that the null hypothesis is true**.

One-sided vs. two-sided hypothesis testing

- The alternative hypothesis $H_A : \mu < 98.6$ or $H_A : \mu > 98.6$ are called **one-sided** alternatives.
- In contrast, the alternative hypothesis $H_A : \mu \neq 98.6$ is **two-sided**.
- For the above three alternatives, the null hypothesis is the same, $H_0 : \mu = 98.6$
- In these cases, the null distribution has a t-distribution; therefore, the hypothesis test is called the t-test.

Hypothesis testing for relationships

Objective

- We now discuss hypothesis testing regarding possible relationships between two variables.
- We focus on problems where we are investigating the relationship between one **binary** categorical variable (e.g., gender) and one **numerical** variable (e.g., body temperature).
- In these situations, the binary variable typically represents two different groups or two different experimental conditions.
- We treat the binary variable (a.k.a., factor) as the explanatory variable in our analysis.
- The numerical variable, on the other hand, is regarded as the response (target) variable (e.g., body temperature).

Relationship Between a Numerical Variable and a Binary Variable

- In general, we can denote the means of the two groups as μ_1 and μ_2.
- The null hypothesis indicates that the population means are equal, $H_0 : \mu_1 = \mu_2$.
- In contrast, the alternative hypothesis is one the following:

$H_A : \mu_1 > \mu_2$ if we believe the mean for group 1 is greater than the mean for group 2.

$H_A : \mu_1 < \mu_2$ if we believe the mean for group 1 is less than the mean for group 2.

$H_A : \mu_1 \neq \mu_2$ if we believe the means are different but we do not specify which one is greater.

Relationship Between a Numerical Variable and a Binary Variable

- We can also express these hypotheses in terms of the **difference** in the means: $H_A : \mu_1 - \mu_2 > 0$, $H_A : \mu_1 - \mu_2 < 0$, or $H_A : \mu_1 - \mu_2 \neq 0$.
- Then the corresponding null hypothesis is that there is no difference in the population means, $H_0 : \mu_1 - \mu_2 = 0$.
- We use a two-sample *t*-test to examine our hypothesis.

Paired t-test

- While we hope that the two samples taken from the population are comparable except for the characteristic that defines the grouping, this is not guaranteed in general.
- To mitigate the influence of other important factors (e.g., age) that are not the focus of our study, we sometimes **pair** (match) each individual in one group with an individual in the other group so that the paired individuals are very similar to each other except for the characteristic that defines the grouping.
- For example, we might recruit twins and assign one of them to the treatment group and the other one to the placebo group.
- Sometimes, the subjects in the two groups are the same individuals under two different conditions.
- When the individuals in the two groups are paired, we use the **paired** t**-test** to take the pairing of the observations between the two groups into account.

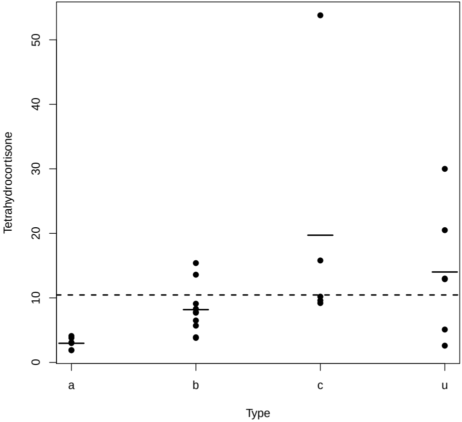
ANOVA

Analysis of Variance

- We discuss Analysis of Variance (ANOVA) models that generalize the t-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.
- The categorical variable is called the **factor** and is typically considered as the explanatory variable.
- In contrast, the numerical variable, whose means across different groups are compared, is regarded as the response variable.

Example

- As an example, we analyze the `Cushings` data set, which is available from the `MASS` package in R.



Variants of ANOVA

- In this example, there was only one factor; this is called *one-way ANOVA*.
- When there are two (or more) factors, we refer to the analysis as *two-way (or multi factor) ANOVA*.
- Sometimes, along with the factors, we have continuous variables, which can explain some of the variation but are not the focus of our analysis; these variables are sometimes called *covariates*.
- The variation of ANOVA that controls for the effect of one (or more) continuous variable (to increase power) is called *Analysis of Covariance (ANCOVA)*.
- It is more common to use linear regression models instead of ANCOVA.

Nonparametric Tests

Wilcoxon rank-sum test

- So far, we discussed parametric tests, where we assume a distribution family for the random variable and express our hypothesis in terms of its parameters.
- Sometimes, we would like to test a hypothesis about a population, but we are reluctant to make any assumption about the distribution of the corresponding random variables (especially if the sample size is small).
- In such cases, we could use nonparametric tests instead.
- Here, we discuss two commonly used nonparametric tests: the **Wilcoxon ranks test** (a.k.a. WilcoxonMannWhitney test) and **Kruskal-Wallis test**.
- These are analogous to the two sample t-test and its generalization, ANOVA, respectively.

Wilcoxon rank-sum test

- Suppose we believe systolic blood pressure is different among people who do not exercise compared to people who exercise regularly (20+ minutes every day).
- Further suppose we have obtained a sample of 12 people, where $m = 5$ of whom do not exercise regularly and $n = 7$ of whom do.
- The observed blood pressure, represented by random variable Y, is given in the following table where observations are sorted by exercise group.

	Non-exercise					Exercise						
Y	140	125	138	124	133	112	121	131	126	115	110	108

Wilcoxon rank-sum test

- To evaluate the null hypothesis, we could examine the **ranks** of the observations.
- The ranks are obtained by arranging the observations in increasing order.
- If the number of observations is $n + m$, the ranks are integers from 1 (for the minimum value) to $n + m$ (for the maximum value).
- In the above example, the ranks are integers from 1 to 12 as shown in the following table:

	Non-exercise					Exercise						
Y	140	125	138	124	133	112	121	131	126	115	110	108
Rank	12	7	11	6	10	3	5	9	8	4	2	1

Wilcoxon rank-sum test

- If blood pressure and exercise are not related, then we should not detect any pattern in the ranks; they should look completely random as if we wrote them without looking at the actual blood pressure values.
- However, if the ranks tend to be lower in one group compared to the other group, then we would regard this as evidence against the null hypothesis that the two distributions are the same and the variables are unrelated.
- To investigate whether the ranks are random, we use the sum of the ranks for one of the two groups as our test statistic.

Kruskal-Wallis test

- The nonparametric equivalent to one-way ANOVA is the **Kruskal-Wallis test**.
- This is a generalization of Wilcoxon rank-sum test and used to compare more than two samples (i.e., the categorical variable identifying groups can take more than two values).
- As an example, we can use this test to analyze the Cushing data discussed above.

Analysis of Categorical Variables

Introduction

- We now discuss **Pearson's χ^2 (chi-squared) test** for testing hypotheses regarding the relationship between two categorical variables.
- Pearson's χ^2 test uses a test statistic, which we denote as Q, to measure the **discrepancy** between the **observed** data and what we **expect to observe under the null** hypothesis (i.e., assuming the null hypothesis is true).
- The null hypothesis in this case states that the two variables are **independent**.
- Recall that for two independent random variables, the joint probability is equal to the product of their individual probabilities.

Smoking and low birthweight babies

- We can create observed and expected contingency tables, and find the observed significance level.

Observed frequency		low	
		0	1
smoke	0	86	29
	1	44	30

Expected frequency		low	
		0	1
smoke	0	79.1	35.9
	1	50.9	23.1

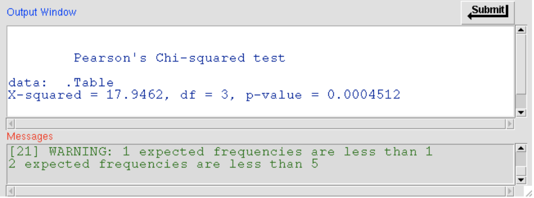
Fisher's exact test

- For Person's χ^2 test to be valid, the expected frequencies (E_{ij}) under the null should be at least 5.
- Occasionally, this requirement is violated (especially when the sample size is small, or the number of categories is large, or some of the categories are rare) and some of the expected frequencies become small (less than 5).
- For example, consider the following table

Frequency		type		
		No	Yes	Total
weight.status	Underweight	2	0	2
	Normal	21	2	23
	Overweight	35	8	43
	Obese	74	58	132
	Total	132	68	200

Fisher's exact test

- If you use Person's χ^2 test, you will receive a warning message:



```
        Pearson's Chi-squared test

data:  .Table
X-squared = 17.9462, df = 3, p-value = 0.0004512
```

```
[21] WARNING: 1 expected frequencies are less than 1
2 expected frequencies are less than 5
```

- To avoid this issue, we use Fisher's exact test.

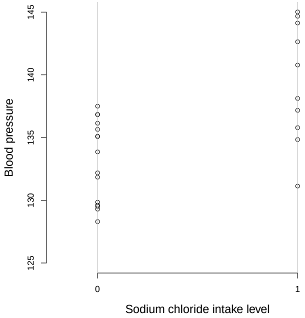
Regression Analysis

Introduction

- We now discuss **linear regression models** for either testing a hypothesis regarding the relationship between one or more **explanatory variables** and a response variable, or **estimating** (predicting) unknown values of the response variable using one or more **predictors**.
- We use X to denote explanatory variables and Y to denote response variables.
- We start by focusing on problems where the explanatory variable is binary. As before, the binary variable X can be either 0 or 1.
- We then continue our discussion for situations where the explanatory variable is numerical.

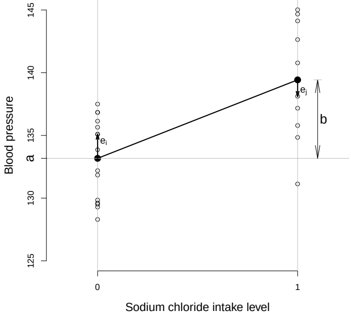
One Binary Explanatory Variable

- Suppose that we want to investigate the relationship between sodium chloride (salt) consumption (low vs. high consumption) and blood pressure among elderly people (e.g., above 65 years old).



One Binary Explanatory Variable

- The following figure shows the dot plot along with sample means, shown as black circles, for each group.
- We connect the two sample means to show the overall pattern for how blood pressure changes from one group to another.



Regression Line

- Using the **intercept** a and **slope** b, we can write the equation for the straight line that connects the estimates of the response variable for different values of X as follows:

$$\hat{y} = a + bx.$$

- The above equation specifies a straight line called the **regression line**.
- The regression line captures the linear relationship between the response variable (here, blood pressure) and the explanatory variable (here, "low" versus "high" sodium chloride diet).

Regression Line

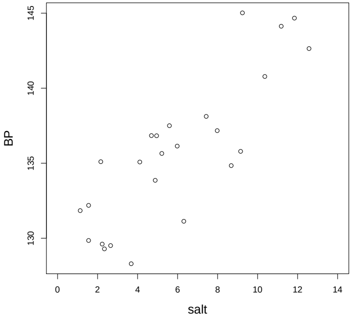
- For this example,

$$\hat{y} = 133.17 + 6.25x.$$

- We **expect** that on average the blood pressure increases by 6.25 units for **one unit increase** in X.
- In this case, one unit increase in X from 0 to 1 means moving from low to high sodium chloride diet group.

One Numerical Explanatory Variable

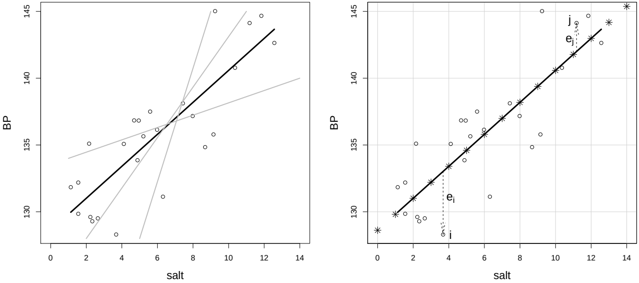
- We now discuss simple linear regression models (i.e., linear regression with only one explanatory variable), where the explanatory variable is numerical.



- As before, we want to find a straight line that captures the relationship between the two variables.

One Numerical Explanatory Variable

- Similar to what we discussed before, among all possible lines we can pass through the data, we choose the **least-squares regression line**, which is the one with the **smallest sum of squared residuals**.



Multiple Linear Regression

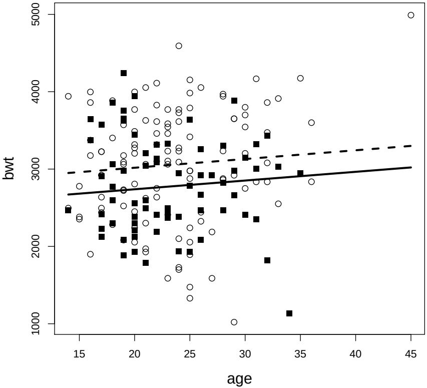


Figure: Nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers

Interaction

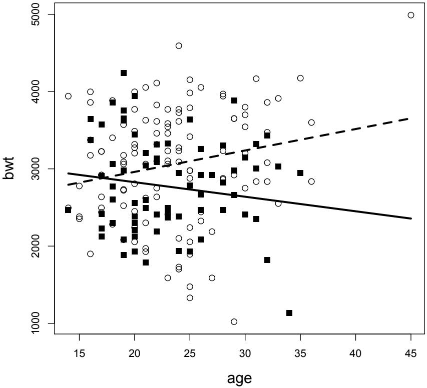


Figure: Nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers

Logistic Regression Models

Introduction

- If the outcome variable is binary, it is common to model the **logit** (i.e., log-odds) of the outcome,

$$\log(\frac{p}{1-p}) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

- Here, p is the probability of the outcome of interest (denoted as 1) given the explanatory variables.
- The resulting models are called **logistic regression models**.

Model evaluation

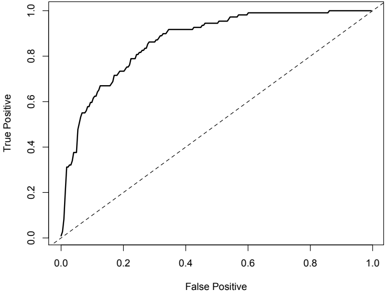
- To evaluate a logistic regression model, we could present the results in a *classification table* as follows:

		Predicted class	
		0	1
True class	0	True Negative	False Positive
	1	False Negative	True Positive

- True positive rate is also called **sensitivity**, and true negative rate is called **specificity**.

ROC curve

- We can also use Receiver Operating Characteristic (ROC) curves, which allow for simultaneous consideration of sensitivity and specificity without setting an arbitrary cutoff.
- The curve plots true positive rate (sensitivity) as a function of false positive rate (1-specificity) for all possible cutoffs from 0 to 1.



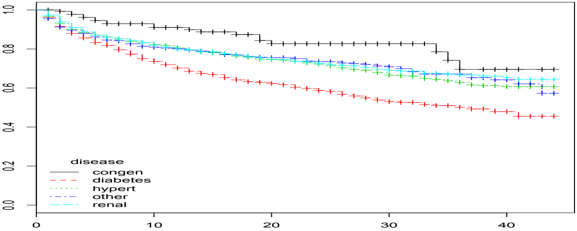
Survival Analysis

Survival analysis

- In survival analysis, the outcome variable is time until a specific event (e.g., death) occurs.
- For each subject, the observed survival time is either time to failure, T, or censoring time, C, whichever comes first.
- An binary event indicator, δ, indicates whether the observed time is time to failure ($\delta = 1$, indicating the event has occurred) or censoring time ($\delta = 0$, indicating that the event occurs after the followup period).
- Survival analysis aims at making inference regarding the time from the origin to the event of interest, and whether the time tends to differ from one group (e.g., treated) to another (e.g., placebo).
- More specifically, we find the survival function, which indicates the probability of survival beyond a certain time t, and compare it across different groups.

Survival analysis

- We usually use the Kaplan-Meier estimator (KM) to estimate survival functions, and use the log-rank test (which is nonparametric) to examine whether the survival distributions are different across two (or more) groups.
- The following plot shows the KM estimates of survival functions using 6805 hemodialysis patients (Sa Carvalho et al., 2003).



Multiple Hypothesis Testing

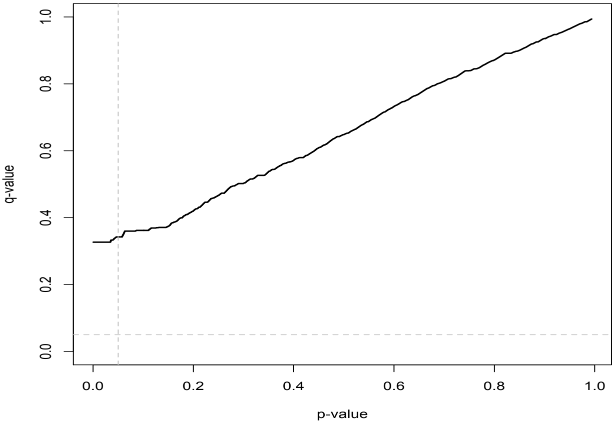
Accounting for multiple hypothesis testing

- So far, we have focused on one hypothesis at a time.
- In recent years, however, high throughput studies with multiple "hypothesis" have become commonplace.
- In such cases, we need to either adjust the p-values or the cutoff used to select a hypothesis as significant.
- A simple procedure to account for multiple hypothesis testing is called Bonferroni correction: given the p-values from m hypotheses, we reject null hypotheses with $p_i \leq \alpha/m$.
- This procedure is however very conservative.

Accounting for multiple hypothesis testing

- Alternatively, we can control the false discovery rate (FDR): the expected proportion of false positive results among reject null hypotheses. For this we find a threshold for p-values such that among hypotheses with p-values below the threshold the expected false positive rate would be at an accepted level (e.g., 0.05).
- Some methods also use FDR-adjusted p-values, which are usually called q-values.

Accounting for multiple hypothesis testing


q-value
p-value