ASTA

The ASTA team

Contents

1 Contingency tables

1.1 A contingency table

- We return to the dataset `popularKids`, where we study **association** between 2 **factors**: `Goals` and `Urban.Rural`.
- Based on a sample we make a cross tabulation of the factors and we get a so-called **contingency table** (krydstabel).

```
popKids <- read.delim("https://asta.math.aau.dk/datasets?file=PopularKids.txt")
library(mosaic)
tab <- tally(~Urban.Rural + Goals, data = popKids, margins = TRUE)
tab
```

```
##             Goals
## Urban.Rural Grades Popular Sports Total
##    Rural        57      50     42   149
##    Suburban     87      42     22   151
##    Urban       103      49     26   178
##    Total       247     141     90   478
```

1.1.1 A conditional distribution

- Another representation of data is the percent-wise distribution of `Goals` for each level of `Urban.Rural`, i.e. the sum in each row of the table is 100 (up to rounding):

```
tab <- tally(~Urban.Rural + Goals, data = popKids)
addmargins(round(100 * prop.table(tab, 1)),margin = 2)
```

```
##             Goals
## Urban.Rural Grades Popular Sports Sum
##    Rural        38      34     28 100
##    Suburban     58      28     15 101
##    Urban        58      28     15 101
```

- Here we will talk about the **conditional distribution** of `Goals` given `Urban.Rural`.
- An important question could be:
    - Are the goals of the kids different when they come from urban, suburban or rural areas? I.e. are the rows in the table significantly different?
- There is (almost) no difference between urban and suburban, but it looks like rural is different.

2 Independence

2.1 Independence

- Recall, that two factors are **independent**, when there is no difference between the population's distributions of one factor given the levels of the other factor.
- Otherwise the factors are said to be **dependent**.
- If we e.g. have the following conditional **population distributions** of `Goals` given `Urban.Rural`:

```
##             Goals
## Urban.Rural Grades Popular Sports
##    Rural       500     300    200
##    Suburban    500     300    200
##    Urban       500     300    200
```

- Then the factors `Goals` and `Urban.Rural` are independent.
- We take a sample and "measure" the factors F_1 and F_2. E.g. `Goals` and `Urban.Rural` for a random child.
- The hypothesis of interest today is:

$$H_0 : F_1 \text{ and } F_2 \text{ are independent}, \quad H_a : F_1 \text{ and } F_2 \text{ are dependent}.$$

2.2 The Chi-squared test for independence

- The relative frequencies in the sample gives an estimate of the unconditional distribution of `Goals`:

```
n <- margin.table(tab)
pctGoals <- round(100 * margin.table(tab, 2)/n, 1)
pctGoals
```

```
## Goals
##  Grades Popular  Sports
##    51.7    29.5    18.8
```

- If we assume independence, then this is also a guess of the conditional distributions of `Goals` given `Urban.Rural`.
- The corresponding expected counts in the sample are then:

```
##              Goals
## Urban.Rural Grades        Popular       Sports        Sum
##    Rural      77.0 (51.7%)  44.0 (29.5%)  28.1 (18.8%) 149.0 (100%)
##    Suburban  78.0 (51.7%)  44.5 (29.5%)  28.4 (18.8%) 151.0 (100%)
##    Urban      92.0 (51.7%)  52.5 (29.5%)  33.5 (18.8%) 178.0 (100%)
##    Sum       247.0 (51.7%) 141.0 (29.5%)  90.0 (18.8%) 478.0 (100%)
```

2.3 Calculation of expected table

```
pctexptab
```

```
##              Goals
## Urban.Rural Grades        Popular       Sports        Sum
##    Rural      77.0 (51.7%)  44.0 (29.5%)  28.1 (18.8%) 149.0 (100%)
##    Suburban  78.0 (51.7%)  44.5 (29.5%)  28.4 (18.8%) 151.0 (100%)
##    Urban      92.0 (51.7%)  52.5 (29.5%)  33.5 (18.8%) 178.0 (100%)
##    Sum       247.0 (51.7%) 141.0 (29.5%)  90.0 (18.8%) 478.0 (100%)
```

- We note that
  - The relative frequency for a given column is columnTotal divided by tableTotal. For example `Grades`, which is $\frac{247}{478} = 51.7\%$.
  - The expected value in a given cell in the table is then the cell's relative column frequency multiplied by the cell's rowTotal. For example `Rural` and `Grades`: $149 \times 51.7\% = 77.0$.
- This can be summarized to:
  - The expected value in a cell is the product of the cell's rowTotal and columnTotal divided by tableTotal.

2.4 Chi-squared (χ^2) test statistic

- We have an **observed table**:

```
tab
```

```
##              Goals
## Urban.Rural Grades Popular Sports
##    Rural         57      50     42
##    Suburban      87      42     22
##    Urban        103      49     26
```

- And an **expected table**, if H_0 is true:

```
##             Goals
## Urban.Rural Grades Popular Sports Sum
##    Rural      77.0    44.0    28.1  149.0
##    Suburban  78.0    44.5    28.4  151.0
##    Urban      92.0    52.5    33.5  178.0
##    Sum       247.0  141.0    90.0  478.0
```

- If these tables are "far from each other", then we reject H_0. We want to measure the distance via the Chi-squared test statistic:
  - $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$: Sum over all cells in the table
  - f_o is the frequency in a cell in the observed table
  - f_e is the corresponding frequency in the expected table.
- We have:

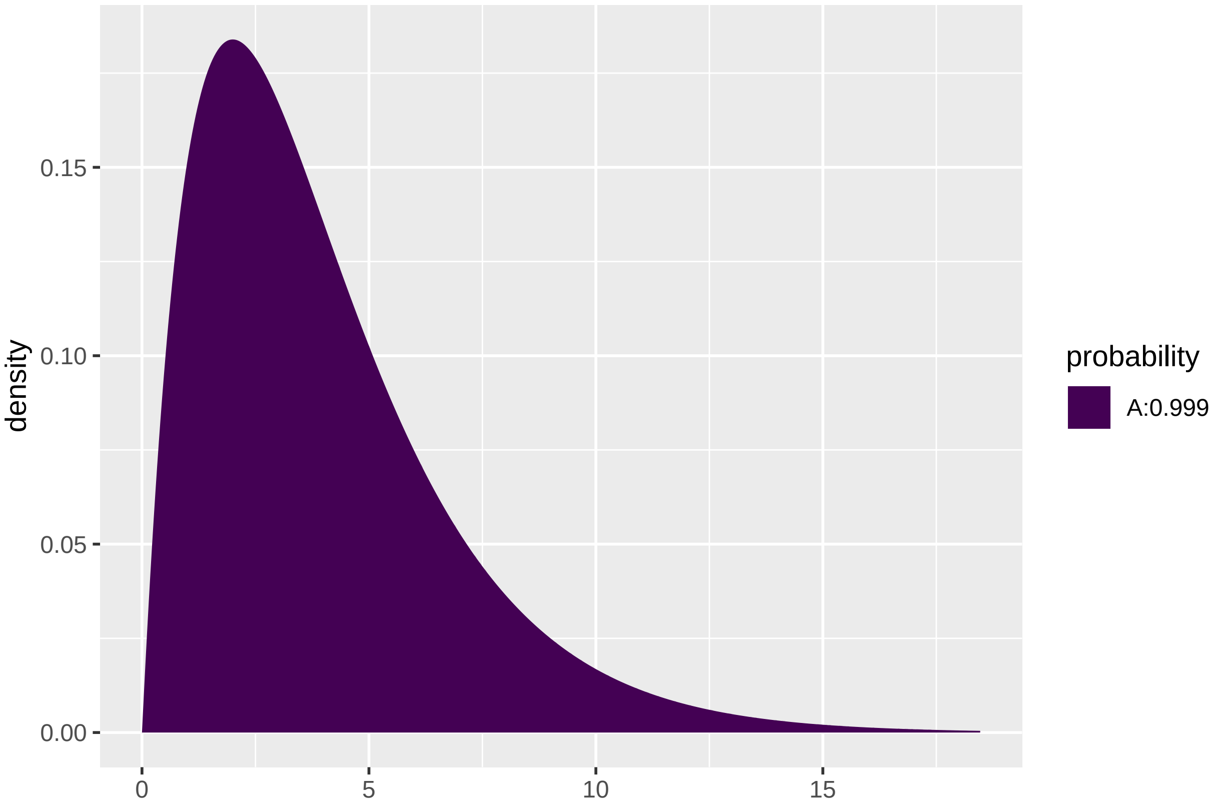
$$X^2_{obs} = \frac{(57-77)^2}{77} + \ldots + \frac{(26-33.5)^2}{33.5} = 18.8$$

- Is this a large distance??

2.5 χ^2-test template.

- We want to test the hypothesis H_0 of independence in a table with r rows and c columns:
  - We take a sample and calculate X^2_{obs} - the observed value of the test statistic.
  - p-value: Assume H_0 is true. What is then the chance of obtaining a larger X^2 than X^2_{obs}, if we repeat the experiment?
- This can be approximated by the χ^2**-distribution** with $df = (r-1)(c-1)$ degrees of freedom.
- For `Goals` and `Urban.Rural` we have $r = c = 3$, i.e. $df = 4$ and $X^2_{obs} = 18.8$, so the p-value is:

```
1 - pdist("chisq", 18.8, df = 4)
```



```
## [1] 0.0008603303
```

- There is clearly a significant association between `Goals` and `Urban.Rural`.

2.6 The function `chisq.test`.

- All of the above calculations can be obtained by the function `chisq.test`.

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 18.828, df = 4, p-value = 0.0008497
```

```
testStat$expected
```

```
##            Goals
## Urban.Rural   Grades  Popular   Sports
##    Rural    76.99372 43.95188 28.05439
##    Suburban 78.02720 44.54184 28.43096
##    Urban    91.97908 52.50628 33.51464
```

- The frequency data can also be put directly into a matrix.

```r
data <- c(57, 87, 103, 50, 42, 49, 42, 22, 26)
tab <- matrix(data, nrow = 3, ncol = 3)
row.names(tab) <- c("Rural", "Suburban", "Urban")
colnames(tab) <- c("Grades", "Popular", "Sports")
tab
```

```
##          Grades Popular Sports
## Rural        57      50     42
## Suburban     87      42     22
## Urban       103      49     26
```

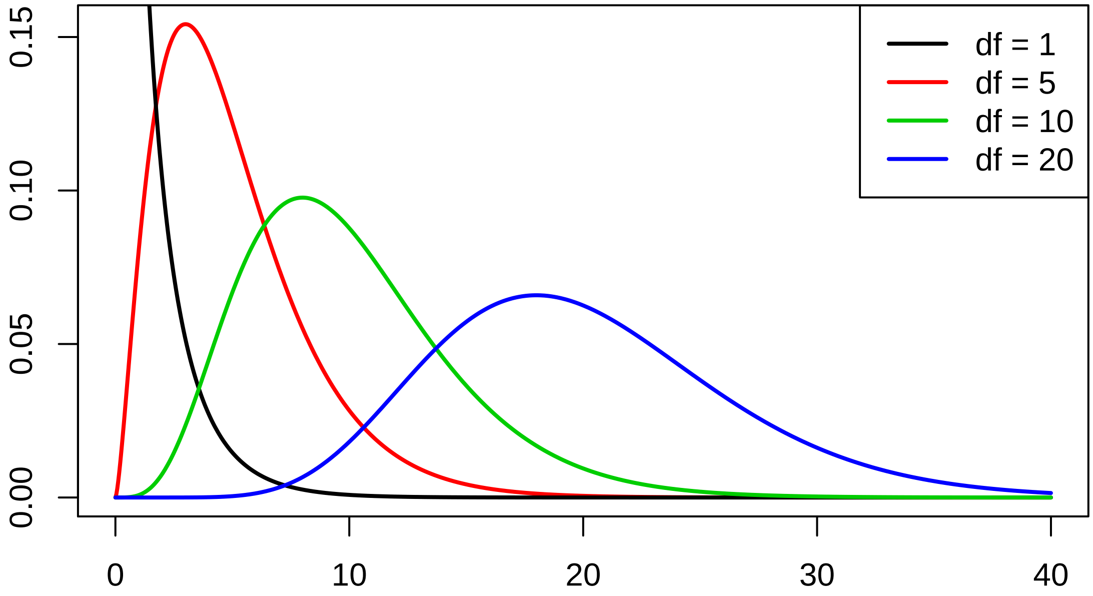
```r
chisq.test(tab)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 18.828, df = 4, p-value = 0.0008497
```

3 The χ^2-distribution

3.1 The χ^2-distribution

- The χ^2-distribution with *df* degrees of freedom:
  - Is never negative.
  - Has mean $\mu = df$
  - Has standard deviation $\sigma = \sqrt{2df}$
  - Is skewed to the right, but approaches a normal distribution when *df* grows.



4 Agresti - Summary

4.1 Summary

- For the the Chi-squared statistic, X^2, to be appropriate we require that the expected values have to be $f_e \geq 5$.
- Now we can summarize the ingredients in the Chi-squared test for independence.

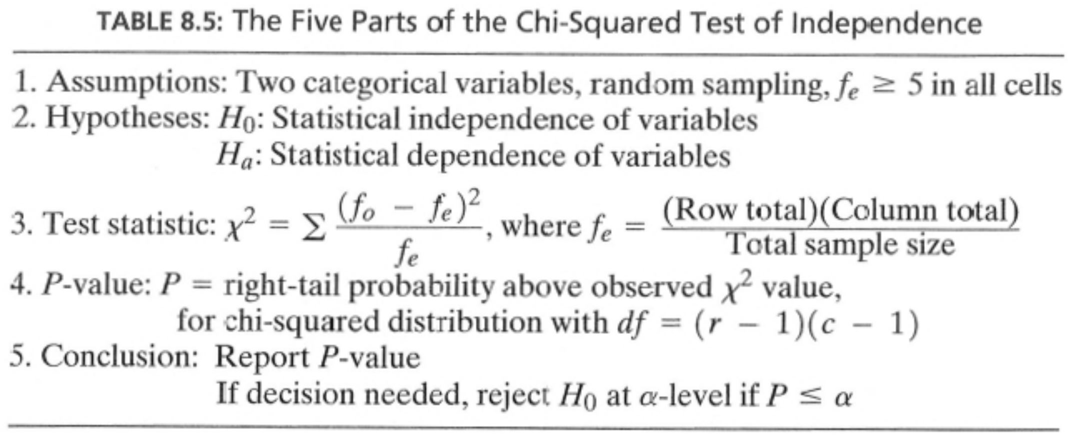


TABLE 8.5: The Five Parts of the Chi-Squared Test of Independence

1. Assumptions: Two categorical variables, random sampling, $f_e \geq 5$ in all cells
2. Hypotheses: H_0: Statistical independence of variables
   H_a: Statistical dependence of variables
3. Test statistic: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, where $f_e = \frac{\text{(Row total)(Column total)}}{\text{Total sample size}}$
4. P-value: P = right-tail probability above observed χ^2 value,
   for chi-squared distribution with $df = (r-1)(c-1)$
5. Conclusion: Report P-value
   If decision needed, reject H_0 at α-level if $P \leq \alpha$

5 Standardized residuals

5.1 Residual analysis

- If we reject the hypothesis of independence it can be of interest to identify the significant deviations.
- In a given cell in the table, $f_o - f_e$ is the deviation between data and the expected values under the null hypothesis.
- We assume that $f_e \geq 5$.
- If H_0 is true, then the standard error of $f_o - f_e$ is given by

$$se = \sqrt{f_e(1 - \text{rowProportion})(1 - \text{columnProportion})}$$

- The corresponding z-score

$$z = \frac{f_o - f_e}{se}$$

  should in 95% of the cells be between ±2. Values above 3 or below -3 should not appear.
- In popKids table cell `Rural and Grade` we got $f_e = 77.0$ and $f_o = 57$. Here columnProportion= 51.7% and rowProportion= $149/478 = 31.2\%$.
- We can then calculate

$$z = \frac{57 - 77}{\sqrt{77(1 - 0.517)(1 - 0.312)}} = -3.95$$

  .
- Compared to the null hypothesis there are way too few rural kids who find grades important.
- In summary: The standardized residuals allow for cell-by-cell (f_e vs f_o) comparision.

5.2 Residual analysis in R

- In R we can extract the standardized residuals from the output of `chisq.test`:

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat$stdres
```

```
##             Goals
## Urban.Rural     Grades    Popular     Sports
##    Rural    -3.9508449  1.3096235  3.5225004
##    Suburban  1.7666608 -0.5484075 -1.6185210
##    Urban     2.0865780 -0.7274327 -1.8186224
```

5.3 Why not just use two-way ANOVA ?

- number of persons in different categories are *not* normally distributed
- variance typically larger the larger expected frequency
- underlying data are discrete (for each person, which column and row category does person belong to)
- these discrete variables are naturally modelled in terms of probabilies for different categories
- therefore hypothesis of independence becomes natural null hypothesis
- it is possible to model table frequencies as dependent variable using a regression model but then we need the framework of *generalized linear models* (see last slides)

Contingency table:

- *counts* of how many individuals fall within different categories for two (or more) categorical variables

Two-way ANOVA:

- a number of individuals/objects/… available for each combination of two categorical variables
- next a continuous variable is measured for each individual or object (this becomes the response variable)

6 Models for table data in R

6.1 Example

- We will study the dataset `HairEyeColor`.

```
HairEyeColor <- read.delim("https://asta.math.aau.dk/datasets?file=HairEyeColor.txt")
head(HairEyeColor)
```

```
##    Hair   Eye  Sex Freq
## 1 Black Brown Male   32
## 2 Brown Brown Male   53
## 3   Red Brown Male   10
## 4 Blond Brown Male    3
## 5 Black  Blue Male   11
## 6 Brown  Blue Male   50
```

- Data is organized such that the variable `Freq` gives the frequency of each combination of the factors `Hair`, `Eye` and `Sex`.
- For example: 32 observations are men with black hair and brown eyes.
- We are interested in the association between eye color and hair color ignoring the sex
- We aggregate data, so we have a table with frequencies for each combination of `Hair` and `Eye`.

```
HairEye <- aggregate(Freq ~ Eye + Hair, FUN = sum, data = HairEyeColor)
HairEye
```

```
##      Eye  Hair Freq
## 1   Blue Black   20
## 2  Brown Black   68
## 3  Green Black    5
## 4  Hazel Black   15
## 5   Blue Blond   94
## 6  Brown Blond    7
## 7  Green Blond   16
## 8  Hazel Blond   10
## 9   Blue Brown   84
## 10 Brown Brown  119
## 11 Green Brown   29
## 12 Hazel Brown   54
## 13  Blue   Red   17
## 14 Brown   Red   26
## 15 Green   Red   14
## 16 Hazel   Red   14
```

6.2 Model specification

- We can write down a model for (the logarithm of) the expected frequencies by using dummy variables z_{e1}, z_{e2}, z_{e3} and z_{h1}, z_{h2}, z_{h3}
- To denote the different levels of `Eye` and `Hair` (the reference level has all dummy variables equal to 0):

$$\log(f_e) = \alpha + \beta_{e1}z_{e1} + \beta_{e2}z_{e2} + \beta_{e3}z_{e3} + \beta_{h1}z_{h1} + \beta_{h2}z_{h2} + \beta_{h3}z_{h3}.$$

- Note that we haven't included an interaction term, which is this case implies, that we assume independence between `Eye` and `Hair` in the model.
- Since our response variable now is a count it is no longer a linear model (`lm`) as we have been used to (linear regression).
- Instead it is a so-called generalized linear model and the relevant `R` command is `glm`.

6.3 Model specification in R

```
model <- glm(Freq ~ Hair + Eye, family = poisson, data = HairEye)
```

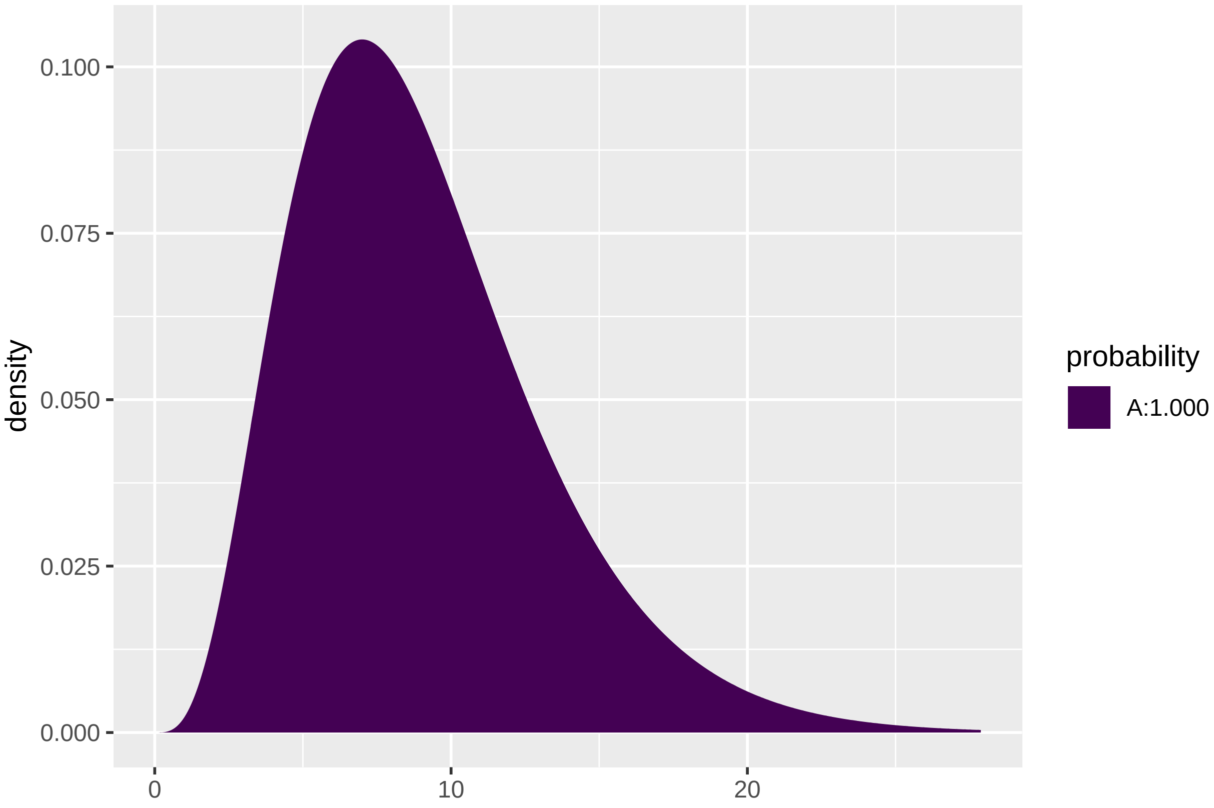
- The argument `family = poisson` ensures that `R` knows that data should be interpreted as discrete counts and not a continuous variable.

```r
summary(model)
```

```
##
## Call:
## glm(formula = Freq ~ Hair + Eye, family = poisson, data = HairEye)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -7.326  -2.065  -0.212   1.235   6.172
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.66926    0.11055  33.191  < 2e-16 ***
## HairBlond    0.16206    0.13089   1.238  0.21569
## HairBrown    0.97386    0.11294   8.623  < 2e-16 ***
## HairRed     -0.41945    0.15279  -2.745  0.00604 **
## EyeBrown     0.02299    0.09590   0.240  0.81054
## EyeGreen    -1.21175    0.14239  -8.510  < 2e-16 ***
## EyeHazel    -0.83804    0.12411  -6.752 1.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 453.31  on 15  degrees of freedom
## Residual deviance: 146.44  on  9  degrees of freedom
## AIC: 241.04
##
## Number of Fisher Scoring iterations: 5
```

- A value of $X^2 = 146.44$ with $df = 9$ shows that there is very clear significance and we reject the null hypothesis of independence between hair and eye color.

```r
1 - pdist("chisq", 146.44, df = 9)
```



```
## [1] 0
```

6.4 Expected values and standardized residuals

- We also want to look at expected values and standardized (studentized) residuals.
- The null hypothesis predicts $e^{3.67+0.02} = 40.1$ with brown eyes and black hair, but we have observed 68.
- This is significantly too many, since the standardized residual is 5.86.
- The null hypothesis predicts 47.2 with brown eyes and blond hair, but we have seen 7. This is significantly too few, since the standardized residual is -9.42.

```
HairEye$fitted <- fitted(model)
HairEye$resid <- rstudent(model)
HairEye
```

```
##      Eye  Hair Freq fitted  resid
## 1   Blue Black   20  39.22 -4.492
## 2  Brown Black   68  40.14  5.856
## 3  Green Black    5  11.68 -2.508
## 4  Hazel Black   15  16.97 -0.583
## 5   Blue Blond   94  46.12  9.368
## 6  Brown Blond    7  47.20 -9.423
## 7  Green Blond   16  13.73  0.719
## 8  Hazel Blond   10  19.95 -2.936
## 9   Blue Brown   84 103.87 -3.437
## 10 Brown Brown  119 106.28  2.151
## 11 Green Brown   29  30.92 -0.511
## 12 Hazel Brown   54  44.93  2.023
## 13  Blue   Red   17  25.79 -2.399
## 14 Brown   Red   26  26.39 -0.101
## 15 Green   Red   14   7.68  2.368
## 16 Hazel   Red   14  11.15  0.961
```

7 Introduction to logistic regression

7.1 Binary response

- We consider a binary response y with outcome 1 or 0. This might be a code indicating whether a person is able or unable to perform a given task.
- Furthermore, we are given an explanatory variable x, which is numeric, e.g. age.
- We shall study models for
$$P(y = 1 \,|\, x)$$
i.e. the probability that a person of age x is able to complete the task.
- We shall see methods for determining whether or not age actually influences the probability, i.e. is y independent of x?

7.2 A linear model

$$P(y = 1 \,|\, x) = \alpha + \beta x$$

is simple, but often inappropiate. If β is positive and x sufficiently large, then the probability exceeds 1.

8 Simple logistic regression

8.1 Logistic model

Instead we consider the **odds** that the person is able to complete the task

$$\texttt{Odds}(y = 1 \,|\, x) = \frac{P(y = 1 \,|\, x)}{P(y = 0 \,|\, x)} = \frac{P(y = 1 \,|\, x)}{1 - P(y = 1 \,|\, x)}$$

which can have any positive value.

The logistic model is defined as:

$$\texttt{logit}(P(y = 1 \,|\, x)) = \log(\texttt{Odds}(y = 1 \,|\, x)) = \alpha + \beta x$$

The function $\texttt{logit}(p) = \log(\frac{p}{1-p})$ - i.e. **log of odds** - is termed **the logistic transformation**.

Remark that log odds can be any number, where zero corresponds to $P(y = 1 \,|\, x) = 0.5$. Solving $\alpha + \beta x = 0$ shows that at age $x_0 = -\alpha/\beta$ you have fifty-fifty chance of solving the task.

8.2 Logistic transformation

- The function `logit()` (remember to load `mosaic` first) can be used to calculate the logistic transformation:

```
p <- seq(0.1, 0.9, by = 0.2)
p
```

```
## [1] 0.1 0.3 0.5 0.7 0.9
```

```
l <- logit(p)
l
```

```
## [1] -2.197 -0.847  0.000  0.847  2.197
```

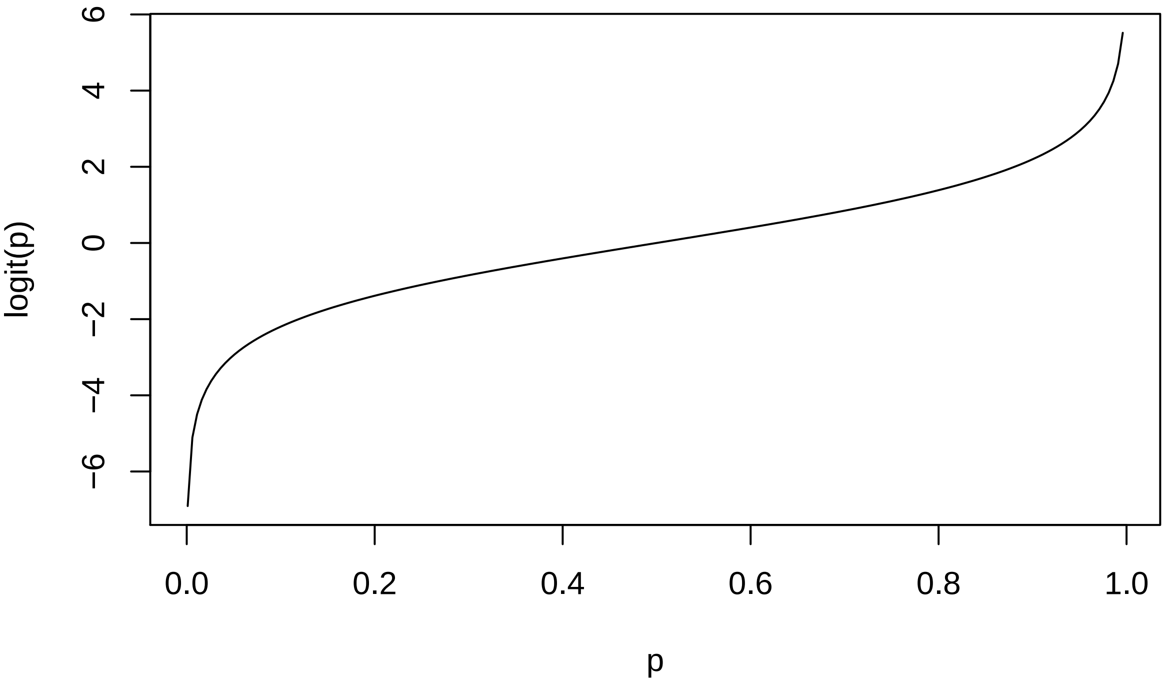
- The inverse logistic transformation `ilogit()` applied to the transformed values can recover the original probabilities:

```
ilogit(l)
```

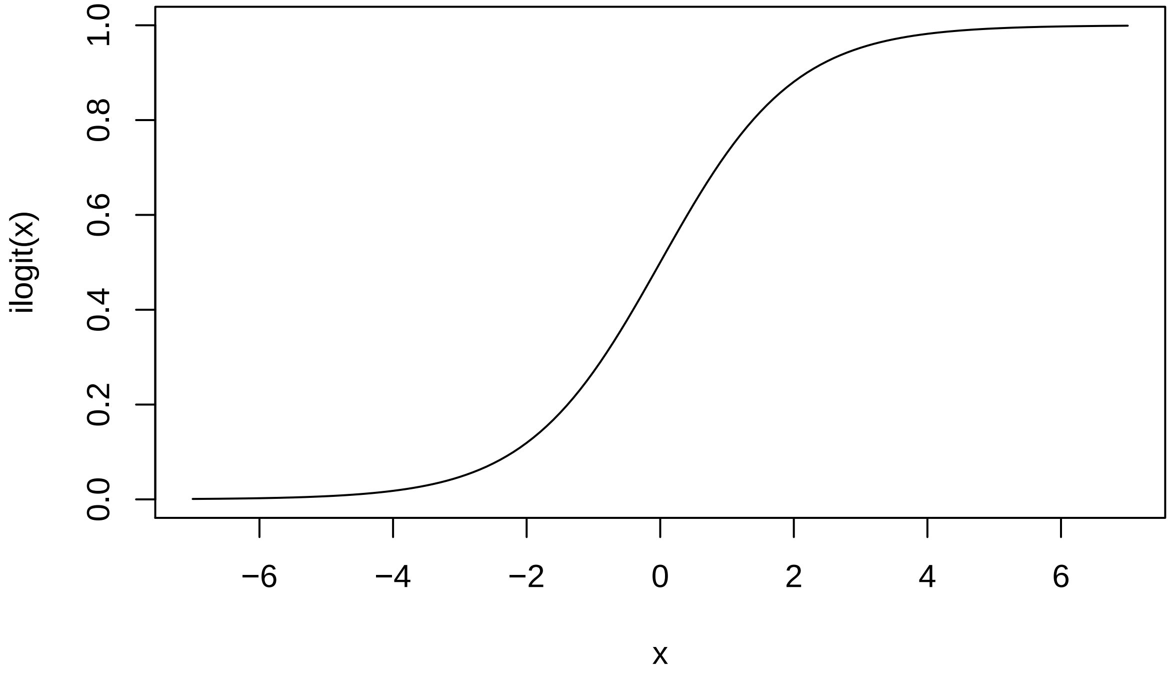
```
## [1] 0.1 0.3 0.5 0.7 0.9
```

Plot of logistic function and inverse logistic

```
p=seq(0.001,0.999,by=0.005)
plot(p,logit(p),type="l")
```



```
x=seq(-7,7,by= 0.1)
plot(x,ilogit(x),type="l")
```



8.3 Odds-ratio

Interpretation of β:

What happens to odds, if we increase age by 1 year?

Consider the so-called **odds-ratio**:

$$\frac{\texttt{Odds}(y=1\,|\,x+1)}{\texttt{Odds}(y=1\,|\,x)} = \frac{\exp(\alpha+\beta(x+1))}{\exp(\alpha+\beta x)} = \exp(\beta)$$

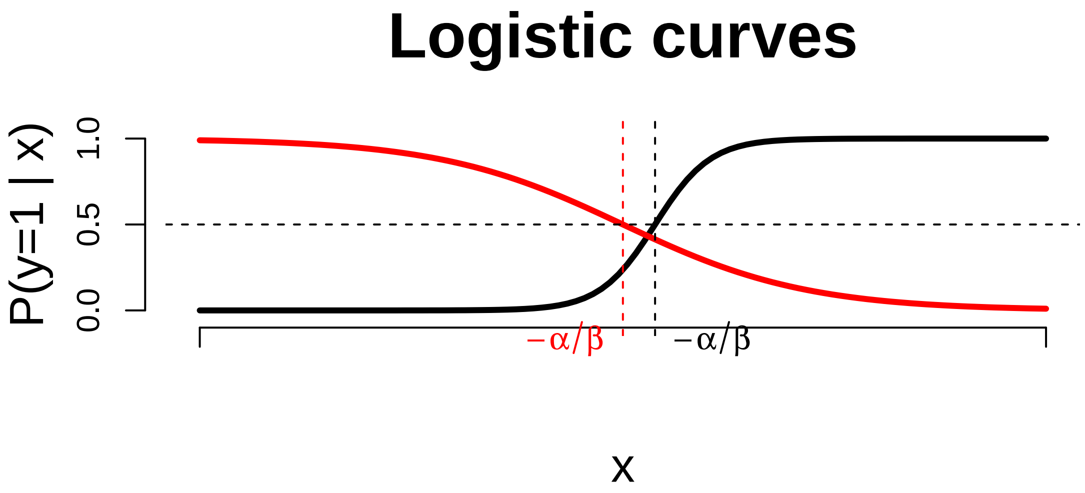
where we see, that $\exp(\beta)$ equals the odds for age $x+1$ relative to odds at age x.

This means that when age increase by 1 year, then the relative change

$$\frac{\exp(\alpha+\beta(x+1)) - \exp(\alpha+\beta x)}{\exp(\alpha+\beta x)}$$

in odds is given by $100(\exp(\beta)-1)\%$.

8.4 Simple logistic regression



Examples of logistic curves for $P(y = 1|x)$. The black curve has a positive β-value (=10), whereas the red has a negative β (=-3).

In addition we note that:

- Increasing the absolute value of β yields a steeper curve.
- When $P(y = 1 \,|\, x) = \frac{1}{2}$ then logit is zero, i.e. $\alpha + \beta x = 0$.

This means that at age $x = -\frac{\alpha}{\beta}$ you have 50% chance to perform the task.

8.5 Example: Credit card data

We shall investigate if income is a good predictor of whether or not you have a credit card.

- Data structure: For each level of income, we let `n` denote the number of persons with that income, and `credit` how many of these that carries a credit card.

```
creInc <- read.csv("https://asta.math.aau.dk/datasets?file=income-credit.csv")
```

```
head(creInc)
```

```
##   Income  n credit
## 1     12  1      0
## 2     13  1      0
## 3     14  8      2
## 4     15 14      2
## 5     16  9      0
## 6     17  8      2
```

8.6 Example: Fitting the model

```
modelFit <- glm(cbind(credit,n-credit) ~ Income, data = creInc, family = binomial)
```

- `cbind` gives a matrix with two column vectors: `credit` and `n-credit`, where the latter is the vector counting the number of persons without a credit card.
- The response has the form `cbind(credit,n-credit)`.
- We need to use the function `glm` (generalized linear model).
- The argument `family=binomial` tells the function that the data has binomial variation. Leaving out this argument will lead `R` to believe that data follows a normal distribution - as with `lm`.
- The function `coef` extracts the coefficients (estimates of parameters) from the model summary:

```
coef(summary(modelFit))
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.518     0.7103   -4.95 7.33e-07
## Income         0.105     0.0262    4.03 5.58e-05
```

8.7 Test of no effect

```
coef(summary(modelFit))
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.518     0.7103   -4.95 7.33e-07
## Income         0.105     0.0262    4.03 5.58e-05
```

Our model for dependence of odds of having a credit card related to income(x) is

$$\texttt{logit}(x) = \alpha + \beta x$$

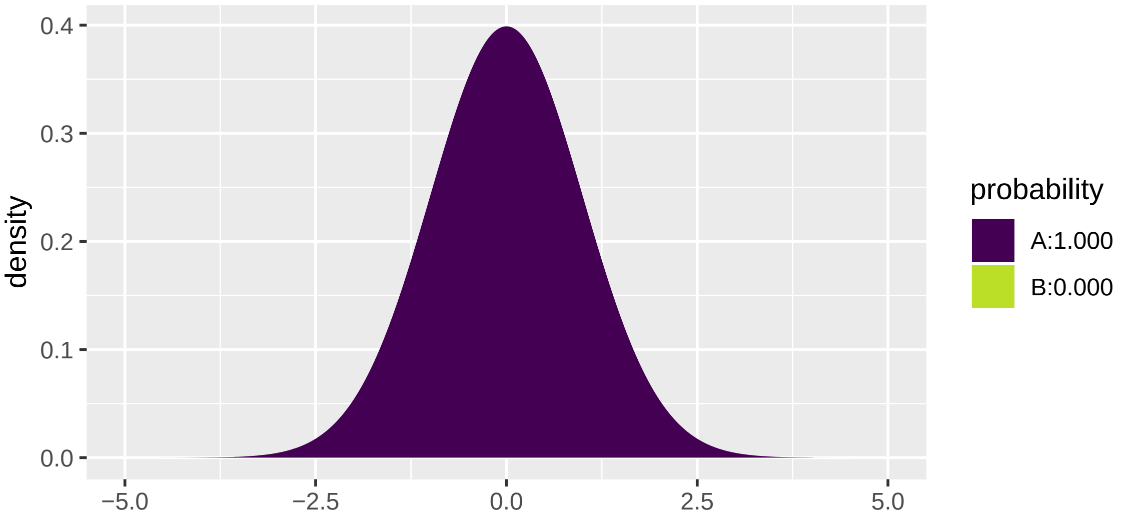
The hypothesis of no relation between income and ability to obtain a credit card corresponds to

$$H_0 : \quad \beta = 0$$

with the alternative $\beta \neq 0$. Inspecting the summary reveals that $\hat{\beta} = 0.1054$ is more than 4 standard errors away from zero.

With a z-score equal to 4.03 we get the tail probability

```
ptail <- 2*(1-pdist("norm",4.03,xlim=c(-5,5)))
```



```
ptail
```

```
## [1] 5.58e-05
```

Which is very significant - as reflected by the p-value.

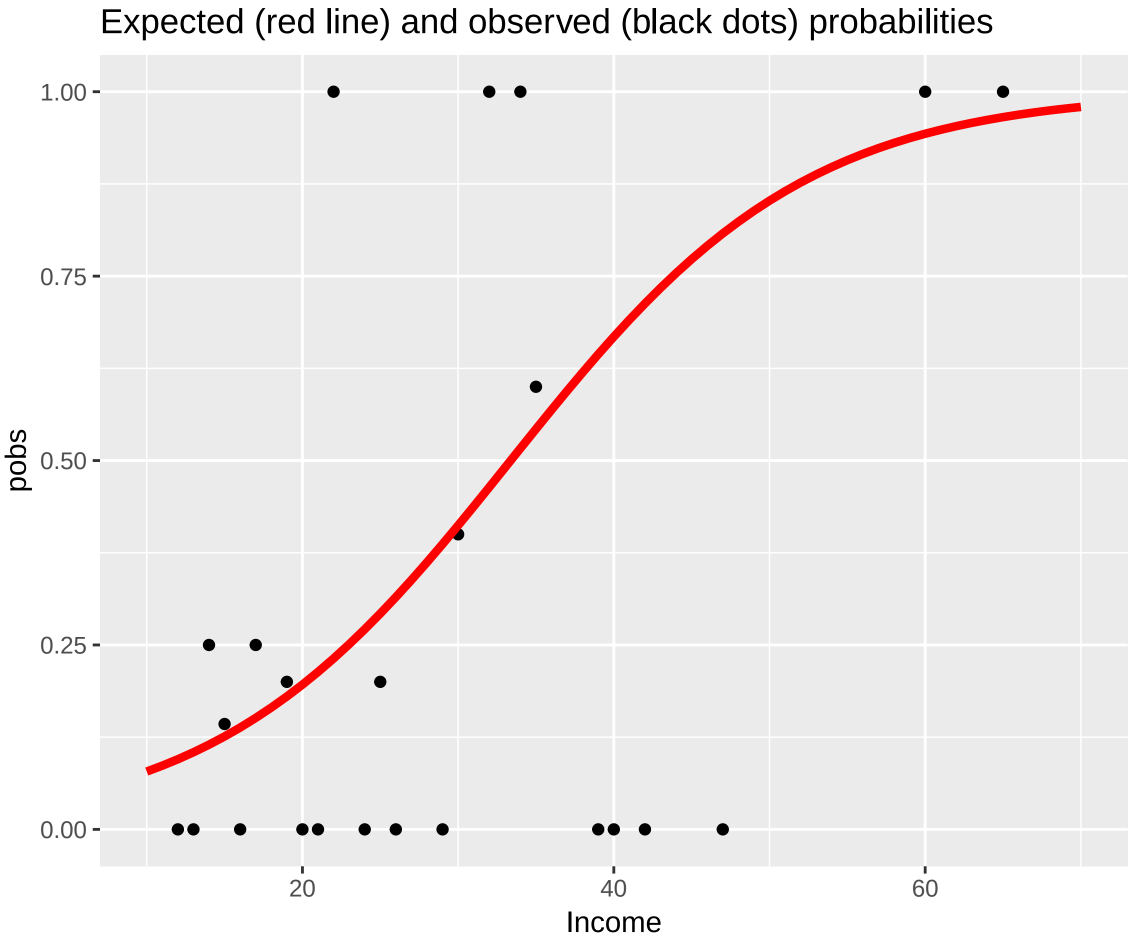
8.8 Confidence interval for odds ratio

From the summary:

- $\hat{\beta} = 0.10541$ and hence $\exp(\hat{\beta}) - 1 = 0.11$. If income increases by 1000 euro, then odds increases by 11%.
- Standard error on $\hat{\beta}$ is 0.02616 and hence a 95% confidence interval for log-odds ratio is $\hat{\beta} \pm 1.96 \times 0.02616 = (0.054; 0, 157)$.
- Corresponding interval for odds ratio: $\exp((0.054; 0, 157)) = (1.056; 1.170)$,

  i.e. the increase in odds is - with confidence 95% - between 5.6% and 17%.

8.9 Plot of model predictions against actual data



- Tendency is fairly clear and very significant.
- Due to low sample size at some income levels, the deviations are quite large.

9 Multiple logistic regression

9.1 Several numeric predictors

We generalize the model to the case, where we have k predictors $x_1, x_2, \ldots, x_k$. Where some might be dummies for a factor.

$$\texttt{logit}(P(y = 1 \mid x_1, x_2, \ldots, x_k)) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Interpretation of β-values is unaltered: If we fix $x_2, \ldots, x_k$ and increase x_1 by one unit, then the relative change in odds is given by $\exp(\beta_1) - 1$.

9.2 Example

Wisconsin Breast Cancer Database covers 683 observations of 10 variables in relation to examining tumors in the breast.

- Nine clinical variables with a score between 0 and 10.
- The binary variable `Class` with levels `benign/malignant`.
- By default `R` orders the levels lexicografically and chooses the first level as reference ($y = 0$). Hence `benign` is reference, and we model odds of `malignant`.

We shall work with only 4 of the predictors, where two of these have been discretized.

```
BC <- read.table("https://asta.math.aau.dk/datasets?file=BC0.dat",header=TRUE)
head(BC)
```

```
##   nuclei cromatin Size.low Size.medium Shape.low     Class
## 1      1        3     TRUE       FALSE      TRUE    benign
## 2     10        3    FALSE        TRUE     FALSE    benign
## 3      2        3     TRUE       FALSE      TRUE    benign
## 4      4        3    FALSE       FALSE     FALSE    benign
## 5      1        3     TRUE       FALSE      TRUE    benign
## 6     10        9    FALSE       FALSE     FALSE malignant
```

9.3 Global test of no effects

First we fit the model `mainEffects` with main effect of all predictors - remember the notation `~ .` for all predictors. Then we fit the model `noEffects` with no predictors.

```
mainEffects <- glm(factor(Class)~., data=BC, family=binomial)
noEffects <- glm(factor(Class)~1, data=BC, family=binomial)
```

First we want to test, whether there is any effect of the predictors, i.e the null hypothesis

$$H_0 : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

9.4 Example

Similarly to `lm` we can use the function `anova` to compare `mainEffects` and `noEffects`. Only difference is that we need to tell the function that the test is a chi-square test and not an F-test.

```
anova(noEffects, mainEffects, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: factor(Class) ~ 1
## Model 2: factor(Class) ~ nuclei + cromatin + Size.low + Size.medium +
##     Shape.low
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       682        884
## 2       677        135  5      749   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`mainEffects` is a much better model.

The test statistic is the Deviance (749.29), which should be small.

It is evaluated in a chi-square with 5 (the number of parameters equal to zero under the nul hypothesis) degrees of freedom.

The 95%-critical value for the $\chi^2(5)$ distribution is 11.07 and the p-value is in practice zero.

9.5 Test of influence of a given predictor

```r
round(coef(summary(mainEffects)),4)
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.709     0.8570  -0.827   0.4080
## nuclei              0.440     0.0823   5.348   0.0000
## cromatin            0.506     0.1444   3.503   0.0005
## Size.lowTRUE       -3.615     0.8081  -4.474   0.0000
## Size.mediumTRUE    -2.377     0.7188  -3.307   0.0009
## Shape.lowTRUE      -2.149     0.6054  -3.550   0.0004
```

For each predictor p can we test the hypothesis:

$$H_0: \quad \beta_p = 0$$

- Looking at the `z`-values, there is a clear effect of all 5 predictors. Which of course is also supported by the p-values.

9.6 Prediction and classification

```r
BC$pred <- round(predict(mainEffects,type="response"),3)
```

- We add the column `pred` to our dataframe `BC`.
- `pred` is the final model's estimate of the probability of `malignant`.

```r
head(BC[,c("Class","pred")])
```

```
##       Class  pred
## 1    benign 0.011
## 2    benign 0.945
## 3    benign 0.017
## 4    benign 0.929
## 5    benign 0.011
## 6 malignant 1.000
```

Not good for patients 2 and 4.

We may classify by `round(BC$pred)`:

- 0 to denote `benign` (probability `BC$pred` less than 0.5)
- 1 to denote `malignant` (probability `BC$pred` more than 0.5)

```
tally(~ Class + round(pred), data = BC)
```

```
##            round(pred)
## Class         0   1
##   benign    433  11
##   malignant  11 228
```

22 patients are misclassified.

```
sort(BC$pred[BC$Class=="malignant"])[1:5]
```

```
## [1] 0.035 0.037 0.089 0.190 0.205
```

There is a malignant woman with a predicted probability of malignancy, which is only 3.5%.

If we assign all women with predicted probability of malignancy above 5% to further investigation, then we only miss two malignant.

```
tally(~ Class + I(pred>.05), data = BC)
```

```
##            I(pred > 0.05)
## Class       TRUE FALSE
##   benign      50   394
##   malignant  237     2
```

The expense is that the number of false positive increases from 11 to 50.

```
tally(~ Class + I(pred>.1), data = BC)
```

```
##            I(pred > 0.1)
## Class       TRUE FALSE
##   benign      27   417
##   malignant  236     3
```

- If we instead set the alarm to 10%, then the number of false positives decreases from 50 to 27.
- But at the expense of 3 false negative.