Categorical Data Analysis - Introduction

Alessio Crippa

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Categorical Data Analysis

Inference on one proportion

2 x 2 Table

Intro Logistic regression

Categorical Data Analysis

Categorical Data Analysis

The analysis of a *response* (or *outcome*) variable that is *categorical*, i.e. has a measurement scale consisiting of categories.

Examples in health sciences: surgery outcome (success, failure), mortality (dead, alive), severity of a disease (none, mild, moderate, severe), ...

We will consider models for the more common cases where the response variable is binary (i.e. can only assume two values 0, 1).

Aim and methods

Describe the binary response (or dependent variable), and possibly, how its distribution changes according to levels of explanatory variables (or independent predictors).

Different methods:

1. univariate analysis and analyses of association (tables)
2. regression models for binary data (logistic and logbinomial)

Marathon data

Dataset: marathon.RData

Outcome: Hyponatremia risk (Serum sodium concentration $\leq$ 135 mmol/liter)

Descriptive abstract

Hyponatremia has emerged as an important cause of race-related death and life-threatening illness among marathon runners. We studied a cohort of marathon runners to estimate the incidence of hyponatremia and to identify the principal risk factors.

Hyponatremia among Runners in the Boston Marathon, New England Journal of Medicine, 2005, Volume 352:1550-1556.

```
load(url("http://alecri.github.io/downloads/data/marathon.Rdata"))
glimpse(marathon)
```

```
Observations: 488
Variables: 18
$ id         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ...
$ na         <dbl> 138, 142, 151, 139, 145, 140, 142, 140, 141, 13...
$ nas135     <fct> na > 135, na > 135, na > 135, na > 135, na > 13...
$ female     <fct> female, male, male, male, female, female, male,...
$ age        <dbl> 24.20534, 44.28200, 41.96304, 28.19713, 30.1820...
$ urinat3p   <fct> >=3, <3, <3, >=3, <3, <3, <3, <3, <3, <3, <3, <...
$ prewt      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ postwt     <dbl> NA, NA, NA, NA, 50.68182, 55.68182, 59.31818, 5...
$ wtdiff     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ height     <dbl> 1.72720, NA, NA, 1.72720, NA, 1.60655, NA, NA, ...
$ bmi        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ runtime    <dbl> NA, 161, 156, 346, 185, 233, 183, 162, 182, 190...
$ trainpace <dbl> 480, 430, 360, 630, NA, NA, 435, 450, 435, 440,...
$ prevmara   <dbl> 3, 40, 40, 1, 3, 25, 19, 2, 80, 10, 16, 3, 2, 8...
$ fluidint   <fct> Every mile, Every mile, Every other mile, Every...
```

Notation

We are interested in making inference on a binary variable Y.
It can only assume two values: 1 with probability p and 0 with probability $1 - p$.

Y is a random variable that follows a *Bernoulli* distribution with parameter p.

$Y \sim \text{Bernoulli}(p)$:
$f_Y(y) = P(Y = y) = p^y(1 - p)^{1-y}$, $y \in \{0, 1\}$.

We want to make inference on the proportion (or probability) of success p.

Inference on one proportion

We randomly sample n observations y_i from a certain population.

OBS: the observations are independent from each other.

We assume that there is underlying population proportion (p) that is the same for all the individuals. We wish to estimate that probability based on our sample.

The OLS method are not particularly appropriate here: the observed values are either 0/1, while the predicted value is a probability. We can instead use an alternative method: **maximum likelihood**.

The likelihood function

The probability for an individual i $P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i}$, $y_i = 0$ or $y_i = 1$.

The probability of observing the sample we have selected is given by the product of individual probabilities:

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = \mathcal{L}(p|(y_1, \ldots, y_n))$$

This quantity is also known as the *likelihood function*. It is a function of the known parameter p, and the sample observations $y_1, \ldots, y_n$.

The method of maximum likelihood provides an estimate of p that maximize the likelihood function, i.e. the probability of obtaining the observed data.

Mathematically and computationally easier to work with the logarithms.

$\ell(p|y_1, \ldots, y_n) = \sum_{i=1}^{n} (y_i \log(p) + (1 - y_i) \log(1 - p))$

Maximum likelihood

Given the data, we maximize the log-likelihood function with respect to p.

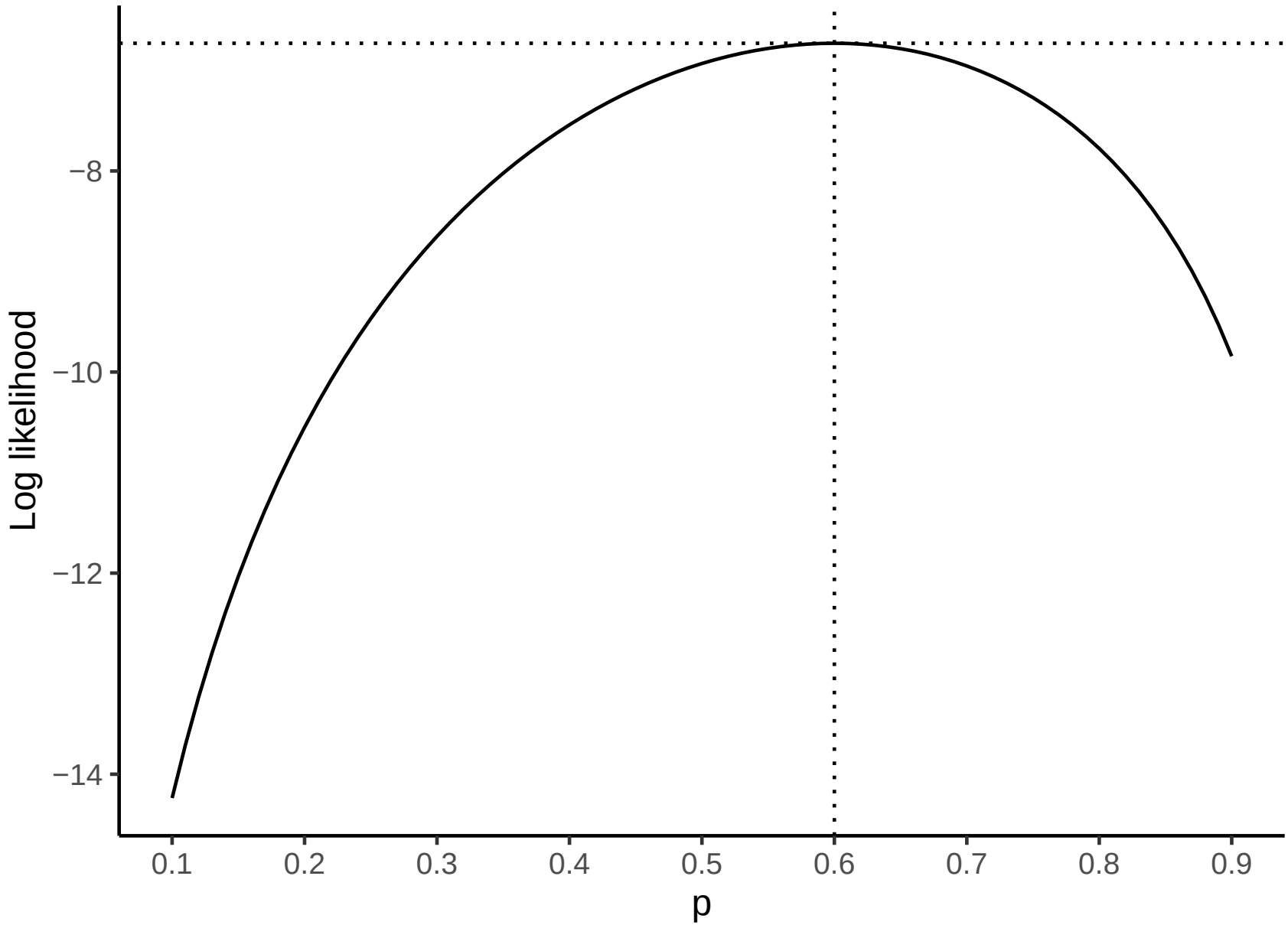
Maximization typically requires an iterative algorithm.

Suppose we collect the following sample of 10 observations: 0, 1, 1, 1, 1, 1, 0, 0, 1, 0

We need to find the value of p that maximizes the likelihood of the observed data.

p	lik	loglik	min_loglik
0.1	0.0000007	-14.237	14.237
0.2	0.0000262	-10.549	10.549
0.3	0.0001750	-8.651	8.651
0.4	0.0005308	-7.541	7.541
0.5	0.0009766	-6.931	6.931
0.6	0.0011944	-6.730	6.730
0.7	0.0009530	-6.956	6.956
0.8	0.0004194	-7.777	7.777
0.9	0.0000531	-9.843	9.843

$\hat{p} = 0.6$ is the more likely value among the considered values.


Log likelihood
−8
−10
−12
−14
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
p

Central Limit Theorem

The maximum likelihood estimate of the population proportion is the sample proportion

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Suppose that we draw 500 observations and 50 of them experienced the outcome of interest ($Y = 1$).

The population proportion p is estimated to be $50/500 = 0.1$ (10%).

Based on a large sample size $n >> 0$, $\hat{P} \sim N(p, \sigma/\sqrt{n})$,

where $\sigma = \sqrt{p(1-p)}$

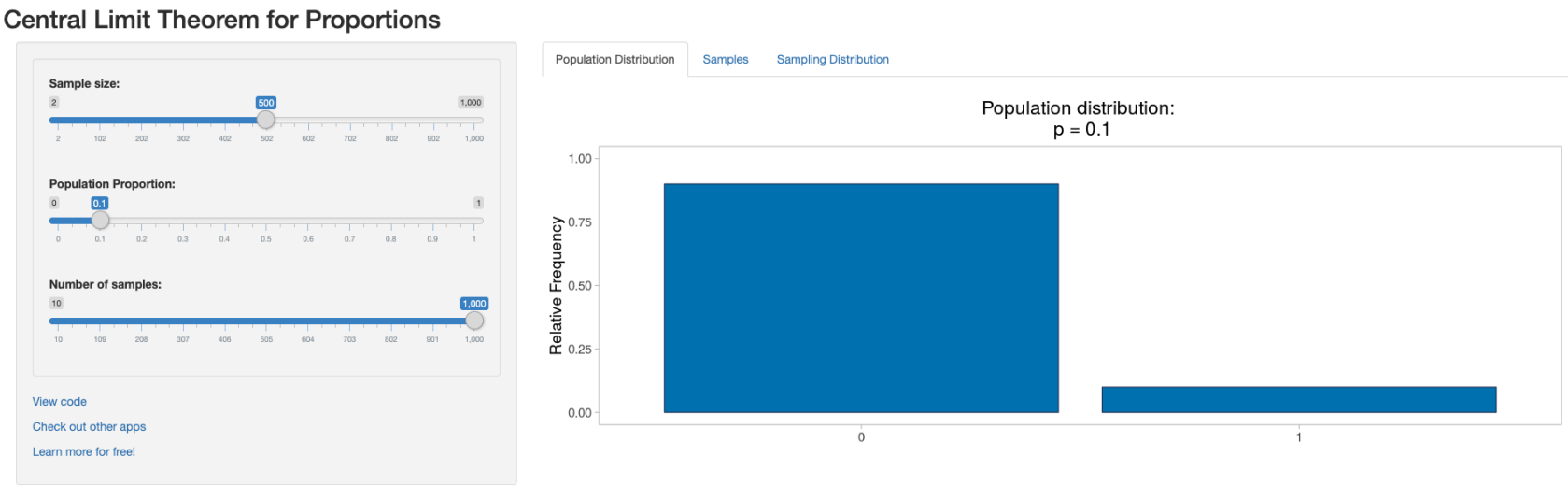
A 95% confidence interval for population proportion p can be obtained as

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

$$0.1 \pm 1.96\sqrt{\frac{0.1(1-0.1)}{500}} = 0.1 \pm 1.96 \cdot 0.013 = (0.07, 0.13)$$

Interactive web app

https://gallery.shinyapps.io/CLT_prop/



Hyponatremia risk

Hyponatremia is a condition that occurs when the level of sodium is very low (in the article defined as 135 mmol per liter or less):

$$Y = \begin{cases} 0 & \text{if na} > 135 \text{ mmol/liter} \\ 1 & \text{if na} \leq 135 \text{ mmol/liter} \end{cases}$$

```
----------------------------
nas135              n    perc
----------------------------
na > 135          426    87.3
na <= 135          62    12.7
----------------------------
```

The estimate for the risk is: $\hat{p} = \frac{62}{426+62} = 0.127$

About thirteen percent of the marathon runners had hyponatremia (a serum sodium concentration of 135 mmol per liter or less).

The 95% confidence interval for the proportion can be calculated as

$$0.127 \pm 1.96\sqrt{0.127(1-0.127)/488} = (0.0975, 0.1565)$$

We are 95% confident that the risk of hyponatremia is between 10% and 16%.

Measures of disease occurrence

- Risk/prevalence: $p = \text{cases}/\text{total} = \sum_{i_1}^{n} y_i/n$
- Odds: $\text{odds} = p/(1-p) = \sum_{i_1}^{n} y_i / \left(n - \sum_{i_1}^{n} y_i\right)$

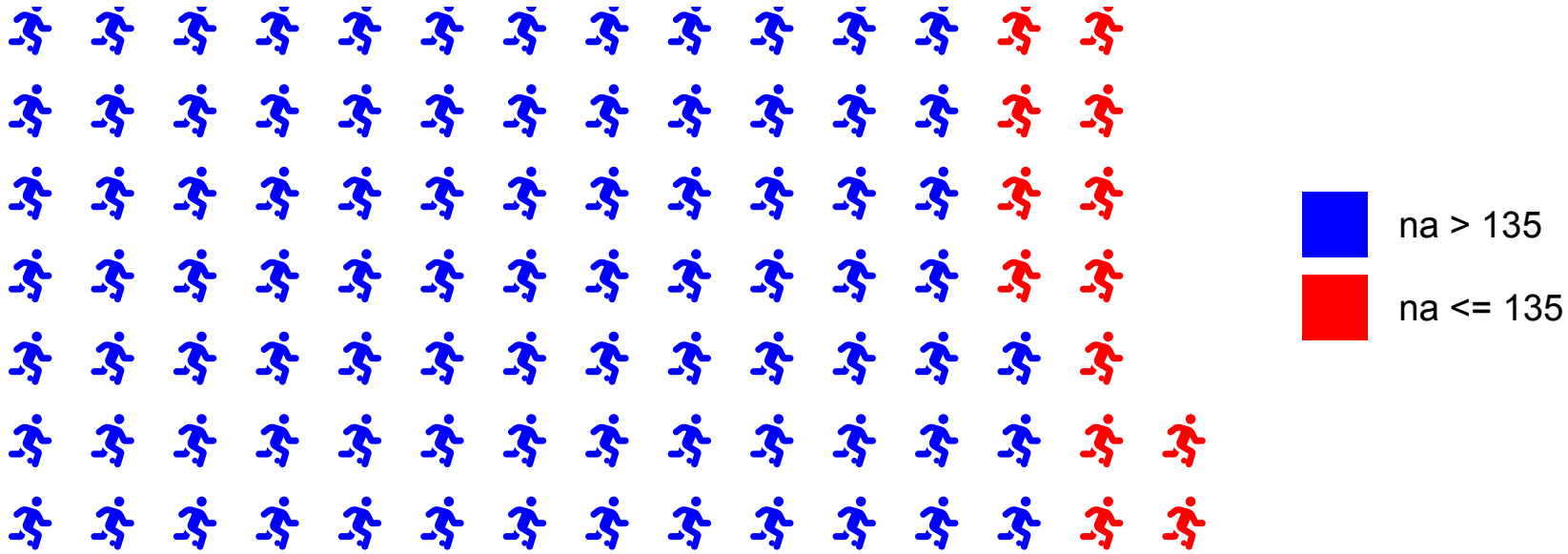
Relations $p = \frac{\text{odds}}{1+\text{odds}}$ and $\text{odds} = p/(1-p)$

The probability or risk of experiencing hyponatremia was $62/488 = 0.13$.

The odds of experiencing hyponatremia was $62/426 = 0.15$.

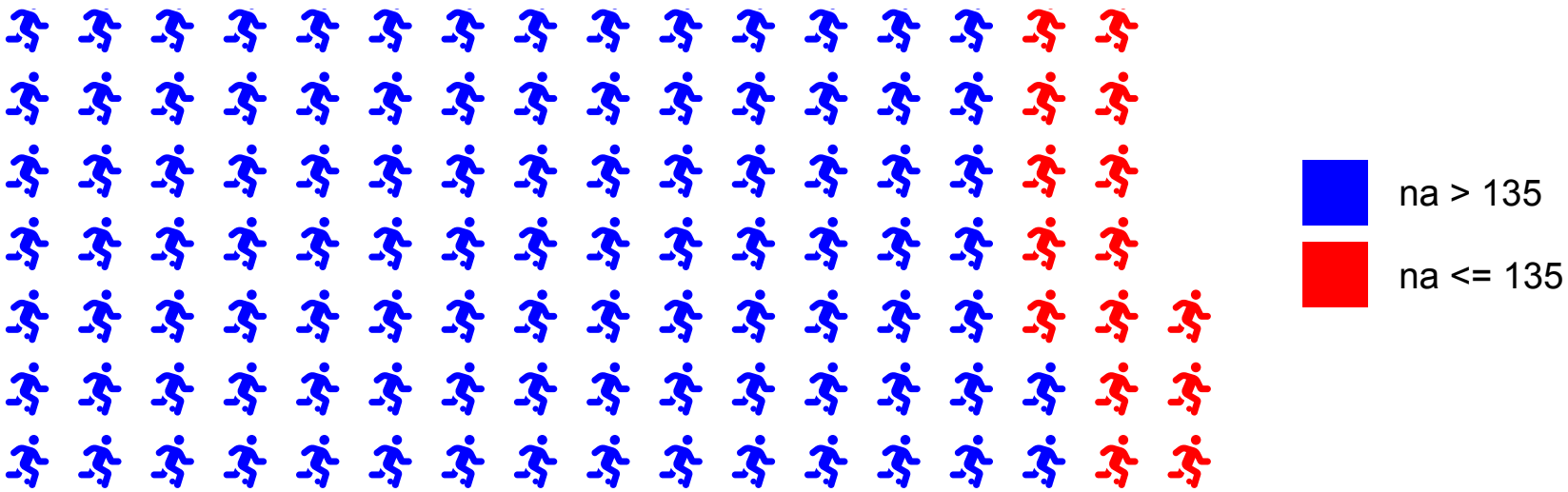
Interpretation of risk

We expect 13 cases for every 100 marathon runners.



Interpretation of odds

We expect 15 cases for every 100 non-cases.



2 x 2 Table

Question: Is the risk of hyponatremia the same for men and women?

	female		
nas135	male	female	Total
na > 135	297	129	426
	92.2	77.7	87.3
na <= 135	25	37	62
	7.8	22.3	12.7
Total	322	166	488
	100.0	100.0	100.0

The Chi-square test

Test if the proportions of cases of hyponatremia is the same for men and women.

$$H_0 : p_1 = p_2$$

The Pearson Chi-Square statistics is based on the comparison of observed (o) and expected (e) counts.

$$X^2 = \sum_{j=1}^{4} \frac{(o_j - e_j)^2}{e_j}$$

Under the null hypothesis of independence $X^2 \sim \chi^2_1$

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|-------------------------|

Total Observations in Table:  488

             |
             |      male |    female | Row Total |
-------------|-----------|-----------|-----------|
    na > 135 |       297 |       129 |       426 |
             |   281.090 |   144.910 |           |
-------------|-----------|-----------|-----------|
   na <= 135 |        25 |        37 |        62 |
             |    40.910 |    21.090 |           |
-------------|-----------|-----------|-----------|
Column Total |       322 |       166 |       488 |
-------------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  20.83654     d.f. =  1     p =  5.001948e-06

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  19.54746     d.f. =  1     p =  9.81313e-06
```

The Pearson Chi-square statistic is 20.84 and the p-value is low (less than 0.01).

We reject the null hypothesis that the proportions of hyponatremia is the same for men and women.

In particular, the risk of hyponatremia among females (22%) is approximately three times the risk of hyponatremia among males (8%).

Measures of association

	Exposed ($X = 1$)	Unexposed ($X = 0$)	
Cases ($Y = 1$)	a	b	m_1
Non-cases ($Y = 0$)	c	d	m_0
	n_1	n_0	n

- Risk Ratio: $RR = \frac{p_1}{p_0}$, with $p_1 = \frac{a}{n_1}$ and $p_0 = \frac{b}{n_0}$
- Odds Ratio: $OR = \frac{a/c}{b/d} = \frac{ad}{bc}$

Relation $RR = OR/(1 - p_0 + p_0 \times OR)$

https://kenkleinman.shinyapps.io/odds-tool/

Confidence interval for Risk Ratio

Inference for RR and OR are based on their log transformation.

$$\log(RR) = \log\left(\frac{p_1}{p_0}\right) = \log(p1) - log(p_0)$$

Based on large sample size ($n >> 0$), the distribution of the sample $\log(RR)$ is approximately normal with mean equal to the population $\log(RR)$ and variance equal to

$$\text{Var}(\log(RR)) = \frac{1}{a} - \frac{1}{n_1} + \frac{1}{b} - \frac{1}{n_0}$$

$$SE(\log(RR)) = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{b} - \frac{1}{n_0}}$$

The 95% CI for the $\log(RR)$ can be calculated as

$$\log(RR) \pm z_{0.975}\text{SE}(\log(RR))$$

The 95% CI for the RR is obtained by exponentiating the 95% CI for the $\log(RR)$

$$\exp\left(\log(RR) \pm z_{0.975}\text{SE}(\log(RR))\right)$$

Confidence interval for Odds Ratio

$$\log(OR) = \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) = \log(p_1)+\log(1-p_0)-log(p_0)-log(1-p_1)$$

Based on large sample size ($n >> 0$), the distribution of the sample $\log(OR)$ is approximately normal with mean equal to the population $\log(OR)$ and variance equal to

$$\text{Var}(\log(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

$$\text{SE}(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

The 95% CI for the $\log(OR)$ can be calculated as

$$\log(OR) \pm z_{0.975}\text{SE}(\log(OR))$$

The 95% CI for the OR is obtained by exponentiating the 95% CI for the $\log(OR)$

$$\exp\left(\log(OR) \pm z_{0.975}\text{SE}(\log(OR))\right)$$

$$RR = \frac{37 \cdot 322}{25 \cdot 166} = 2.872$$

$$\log(RR) = \log(2.872) = 1.055$$

$$\text{SE}(\log(RR)) = \sqrt{\frac{1}{37} - \frac{1}{166} + \frac{1}{25} - \frac{1}{322}} = 0.241$$

95% CI for $\log(RR)$: $1.055 \pm 1.96 \cdot 0.241 = (0.583, 1.527)$

95% CI for RR: $(\exp(0.583), \exp(1.527)) = (1.791, 4.604)$

Interpretation: the risk of hyponatremia among women is 2.872 (95% CI 1.791, 4.604) **times** higher than in men.

$$OR = \frac{37 \cdot 297}{25 \cdot 129} = 3.408$$

$$\log(OR) = \log(3.408) = 1.226$$

$$\text{SE}(\log(OR)) = \sqrt{\frac{1}{37} + \frac{1}{25} + \frac{1}{129} + \frac{1}{297}} = 0.28$$

95% CI for $\log(OR)$: $1.226 \pm 1.96 \cdot 0.28 = (0.677, 1.775)$

95% CI for OR: $(\exp(0.677), \exp(1.775)) = (1.968, 5.9)$

Interpretation: the odds of hyponatremia among women is 3.408 (95% CI 1.968, 5.9) **times** higher than in men.

```
with(marathon,
     twoby2(exposure = relevel(female, 2), outcome = relevel(nas135, 2)))
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome    : na <= 135
Comparing : female vs. male

       na <= 135 na > 135    P(na <= 135) 95% conf. interval
female        37      129          0.2229     0.166   0.2925
male          25      297          0.0776     0.053   0.1124

                                   95% conf. interval
            Relative Risk: 2.8708    1.7914   4.6007
        Sample Odds Ratio: 3.4074    1.9701   5.8936
Conditional MLE Odds Ratio: 3.3982    1.9037   6.1528
    Probability difference: 0.1453    0.0790   0.2186

             Exact P-value: 0
         Asymptotic P-value: 0
------------------------------------------------------
```

Intro Logistic regression

Why not linear regression?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Major problems:

- it is not appropriate to model Y as a linear function of the parameters because Y has only two values;
- the predicted values can be any positive or negative number, not just 0 or 1;
- the conditional distribution of $Y|X$ is Bernoulli, not normal;
- the values of 0 and 1 are arbitrary.

The important part is not to predict the numerical value of Y, but the (conditional) probability that success or failure occurs.

$$P(Y_i = 1|X_i) = \mathrm{E}[Y_i = 1|X_i] = \beta_0 + \beta_1 X_i$$

Major problems:

- the right hand side of the equation can be any number, but the left hand side can only range from 0 to 1;
- It turns out the relationship is not linear, but rather follows an S-shaped (or sigmoidal) curve.

To obtain a linear relationship, we need to transform this response too, $P(Y_i = 1|X_i)$. The function needs to

- not to be restricted to values between 0 and 1;
- form a linear relationship with our parameters.

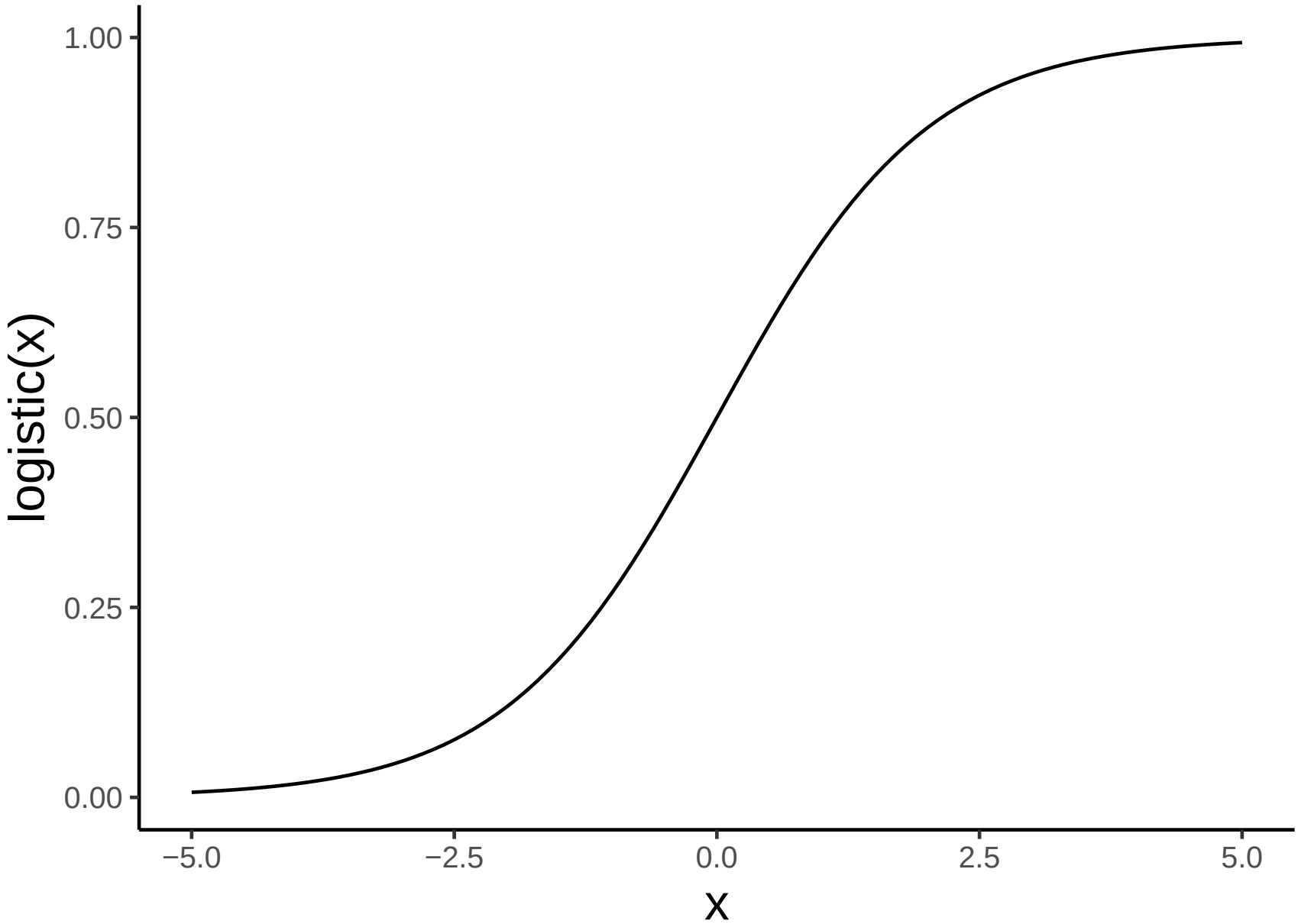
The logistic function

The **logistic function** describes the mathematical form on which the **logistic model** is based.

This function is defined as

$$\text{logistic}(x) = \frac{e^x}{1 + e^x}$$

It describes an S-shape curve


1.00
0.75
0.50
0.25
0.00
logistic(x)
−5.0
−2.5
0.0
2.5
5.0
x

- The logistic function ranges between 0 and 1 and it is probably the main reason the logistic model is so popular.
- The model is designed to describe a probability, which is always some number between 0 and 1.
- In epidemiological terms, such a probability gives the risk of an individual getting a disease.

Logistic regression model

It is mathematical model to make inference on the probability of a binary outcome given a set of covariates.

It can be used for any type of exposure: binary, continuous, or categorical covariate values.

It allows adjustment for confounding, assessment of effect modification (interaction).

Estimation method: Maximum likelihood (yields point estimates, standard error estimates, confidence intervals, and p-values)

The empty model (no predictors)

The logistic model can be defined as

$$\text{logit}\left(P(Y)\right) = \log\left(\frac{P(Y)}{1-P(Y)}\right) = \log\left(\text{odds}(Y)\right) = \beta_0$$

Recalling the relation between odds and probability ($p = \text{odds}/(1+\text{odds})$), the model can also be written as

$$P\left(Y=1\right) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} = \text{invlogit}(\beta_0)$$

The last equation is ofter referred to as the `invlogit` function. It is very useful to go from *log odds* to *risk* of the outcome.

```r
mod0 <- glm(nas135 ~ 1, data = marathon, family = "binomial")
summary(mod0)
```

```
Call:
glm(formula = nas135 ~ 1, family = "binomial", data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5213  -0.5213  -0.5213  -0.5213   2.0313

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9273     0.1359  -14.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 371.6  on 487  degrees of freedom
Residual deviance: 371.6  on 487  degrees of freedom
AIC: 373.6

Number of Fisher Scoring iterations: 4
```

Interpretation of regression coefficient

$\hat{\beta}_0 = -1.93$ is an estimate of the **log odds** of hyponatremia.

The interpretation of the coefficients (on the log scale) is quite cumbersome. It is usually better to interpret their exponential

$\exp(\hat{\beta}_0) = \exp(-1.93) = 0.15$ is an estimate of the **odds** hyponatremia

We expect 15 cases for every 100 non-cases.

We are often interested to express the results in terms of predicted probabilities:

$$\hat{P}(Y = 1) = \text{invlogit}(\log(odds)) = \text{invlogit}(\hat{\beta}_0) = \frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)} = \frac{\exp(-1.93)}{1+\exp(-1.93)} = 0.13$$

We expect 13 cases for every 100 runners.

Confidence interval for predicted probabilities

The β coefficients are estimated by maximum likelihood. For large samples, the distribution of the β coefficients is approximately normal and confidence interval based on standard normal distribution can be constructed.

NB the confidence intervals are constructed on the log odds.

95% CI for the **log** odds:

$$\hat{\beta}_0 \pm 1.96 \cdot \text{SE}(\hat{\beta}_0) = -1.93 \pm 1.96 \cdot 0.136 = (-2.2, -1.66)$$

The 95% CI for $P(Y)$ is the invlogit of the 95% CI of the log odds.

$$\left(\frac{\exp(-2.2)}{1+\exp(-2.2)}, \frac{\exp(-1.66)}{1+\exp(-1.66)} \right) = (0.111, 0.19)$$