# Advanced statistics and modelling

6. week

# Parameter inference in statistics
# Hypothesis testing, MLE, Bootstrap
# Excercises

# Working with sheets



- Raise your hand and remember your serial number!
- Open weblink
  - Open file: ex1 in your browser
- Open weblink in a new window
  - Go to folder: sandbox
  - Open file with your serial number, rename one sheet to random code (eg. your neptun code)
    This indicates for your fellows, that this file is in use, they should not work with it accidentally
- Copy and paste data from sheet "master" to sheet "work"
- Be careful: avoid copying the top left most cell (A1)!
- Now you are ready to work with your sheet: add some formulas, data etc in sheet "work".
- In case you have modified your "master" sheet, it will probably cause an error. Simply "undo" your modification.

# Working with sheets



- Now you are ready to work with your sheet: add some formulas, data etc in sheet "work".
  - In case you have modified your "master" sheet, it will probably cause an error. Simply "undo" your modification.

Note: if your browser is set to a default language other than English, you will see translated version of functions and menus of the instructing screen.
Do not worry, in most cases you can give function names in English in this case as well.

# Method of moments, MME 1.

- Find the appropriate distribution of inter arrival times!
  - Which distribution?
  - Express the first moment with the parameter of the distribution!
- Check the fitted parameter: compare distribution from data and fitted PDF! (Use QQ-plot!)
- Calculate the residuals asw ell!

Note:

$$
\begin{aligned}
F(x) &= 1 - e^{-\lambda x} \\
f(x) &= \lambda e^{-\lambda x} \\
\langle x \rangle &= 1/\lambda
\end{aligned}
$$

# Method of Momements, MME 2.



- Set up a dataset with uniformly distributed random numbers! You can choose to copy numbers from your "master" sheet, or from the "ex1" file (sheet: unif), or generate your self with the function "rand()".
- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.

How are the parameters, the moments and the QQ-plot changing?

# Method of Momements, MME 2.

- Set up a dataset with uniformly distributed random numbers! You can choose to copy numbers from your "master" sheet, or from the "ex1" file (sheet: unif), or generate your self with the function "rand()".
- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.
- Prepare a QQ-plot for testing the fitted values!
  - Fit the best line on the QQ-plot!

How are the parameters, the moments and the QQ-plot changing?

# Method of Momements, MME 2.

- Set up a dataset with uniformly distributed random numbers! You can choose to copy numbers from your "master" sheet, or from the "ex1" file (sheet: unif), or generate your self with the function "rand()".
- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.
- Prepare a QQ-plot for testing the fitted values!
  - Fit the best line on the QQ-plot!
- Experiment with the parameters: try to
  - shift the raw dataset,

How are the parameters, the moments and the QQ-plot changing?

# Method of Momements, MME 2.



- Set up a dataset with uniformly distributed random numbers! You can choose to copy numbers from your "master" sheet, or from the "ex1" file (sheet: unif), or generate your self with the function "rand()".
- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.
- Prepare a QQ-plot for testing the fitted values!
  - Fit the best line on the QQ-plot!
- Experiment with the parameters: try to
  - shift the raw dataset,
  - rescale the dataset,

How are the parameters, the moments and the QQ-plot changing?

# Method of Momements, MME 2.



- Set up a dataset with uniformly distributed random numbers! You can choose to copy numbers from your "master" sheet, or from the "ex1" file (sheet: unif), or generate your self with the function "rand()".
- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.
- Prepare a QQ-plot for testing the fitted values!
  - Fit the best line on the QQ-plot!
- Experiment with the parameters: try to
  - shift the raw dataset,
  - rescale the dataset,
  - insert outliers into the data

How are the parameters, the moments and the QQ-plot changing?

# Method of Momements, MME 2.



- Try to estimate the parameters of the uniform distribution for your dataset!
  - Note: you have to find two parameters now, so you need to calculate two moments.
- Prepare a QQ-plot for testing the fitted values!
  - Fit the best line on the QQ-plot!
- Experiment with the parameters: try to
  - shift the raw dataset,
  - rescale the dataset,
  - insert outliers into the data

How are the parameters, the moments and the QQ-plot changing?
Note: in real scenarios you will meet similar biased or transformed data usually.

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?
- Calls can be modeled by a stationary Poisson process. How many parameters do you need to fit?

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?
- Calls can be modeled by a stationary Poisson process. How many parameters do you need to fit?
- Prepare a QQ-plot for testing the fitted values! Take into account, that you work with a discrete distribution!

Poisson:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?
- Calls can be modeled by a stationary Poisson process. How many parameters do you need to fit?
- Prepare a QQ-plot for testing the fitted values! Take into account, that you work with a discrete distribution!
- Fit the best line on the QQ-plot! Check the slope and the abscissa of the line! Correspond these numbers to your expectations?

Poisson:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?
- Calls can be modeled by a stationary Poisson process. How many parameters do you need to fit?
- Prepare a QQ-plot for testing the fitted values! Take into account, that you work with a discrete distribution!
- Fit the best line on the QQ-plot! Check the slope and the abscissa of the line! Correspond these numbers to your expectations?
- Plot the residuals and plot the cumulative distribution functions! Can you retain the Poissonian assumption?

Poisson:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

# Method of Moments, MME 3



- We have a dataset with counts of incoming calls at the secretary of the dean for each working hours in a week. Assuming constant calling rate, try to fit an appropriate distribution for the number of calls! (sheet: "pp")
  Which distributions would you try first?
- Calls can be modeled by a stationary Poisson process. How many parameters do you need to fit?
- Prepare a QQ-plot for testing the fitted values! Take into account, that you work with a discrete distribution!
- Fit the best line on the QQ-plot! Check the slope and the abscissa of the line! Correspond these numbers to your expectations?
- Plot the residuals and plot the cumulative distribution functions! Can you retain the Poissonian assumption?

Conclusions:QQ-plot indicates, that this is the good distribution family. But further parameters are to be fitted: inhomogeneous Poisson process could be better.
Poisson:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

# Maximum Likelihood Estimation, MLE



- Find the MLE for the uniform distribution using the data from the MME example! (sheet: "unif")

# Maximum Likelihood Estimation, MLE



- Find the MLE for the uniform distribution using the data from the MME example! (sheet: "unif")
- Recall: PDF of uniform

$$f(x) = \frac{1}{b-a} \quad x \in [a, b]$$

- Can you solve it? How do you interpret the result?

# Maximum Likelihood Estimation, MLE



- Find the MLE for the uniform distribution using the data from the MME example! (sheet: "unif")
- Recall: PDF of uniform

$$f(x) = \frac{1}{b-a} \quad x \in [a, b]$$

- MLE:

$$\mathcal{L}(a, b) = \prod_i f(x_i; a, b) = \frac{1}{(b-a)^n}$$

  where $x_i$ are fixed numbers from the $n$ data.
- Can you solve it? How do you interpret the result?

# Maximum Likelihood Estimation, MLE



- Find the MLE for the uniform distribution using the data from the MME example! (sheet: "unif")
- Recall: PDF of uniform

$$f(x) = \frac{1}{b-a} \quad x \in [a, b]$$

- MLE:

$$\mathcal{L}(a, b) = \prod_i f(x_i; a, b) = \frac{1}{(b-a)^n}$$

  where $x_i$ are fixed numbers from the $n$ data.
- $\mathcal{L}$ seems to be good behaving function, find the maximum with usual analysis:

$$\partial_a \mathcal{L} = 0$$

$$\partial_b \mathcal{L} = 0$$

- Can you solve it? How do you interpret the result?

# Maximum Likelihood Estimation, MLE

- Because $f(x) = 0$ if $x \notin [a,b]$, we have

$$a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$$

where data $(x_i)$ are in increasing order.

# Maximum Likelihood Estimation, MLE



- Because $f(x) = 0$ if $x \notin [a, b]$, we have

$$a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$$

where data $(x_i)$ are in increasing order.

- $\mathcal{L}$ is not differentiable in $x_1$ and $x_n$!

# Maximum Likelihood Estimation, MLE



- Because $f(x) = 0$ if $x \notin [a, b]$, we have

$$a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$$

  where data $(x_i)$ are in increasing order.
- $\mathcal{L}$ is not differentiable in $x_1$ and $x_n$!
- MLE:

$$(a, b) = (min(x), max(x))$$

  But: this is a biased estimation.
  (Recall: what is a biased estimation?)

# Maximum Likelihood Estimation, MLE



- Because $f(x) = 0$ if $x \notin [a, b]$, we have

$$a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$$

  where data $(x_i)$ are in increasing order.

- $\mathcal{L}$ is not differentiable in $x_1$ and $x_n$!
- MLE:

$$(a, b) = (min(x), max(x))$$

  But: this is a biased estimation.
  (Recall: what is a biased estimation?)

- Try to correct the results to have an unbiased estimation with estimation values:

$$(a, b) = (\langle min(x) \rangle, \langle max(x) \rangle)$$

# Maximum Likelihood Estimation, MLE



- Calculate $\langle max(x) \rangle$ from the CDF!

# Maximum Likelihood Estimation, MLE



- Calculate $\langle max(x) \rangle$ from the CDF!
  - Utilize the independence of data points!

# Maximum Likelihood Estimation, MLE



- Calculate $\langle max(x) \rangle$ from the CDF!
- Utilize the independence of data points!

$$
\begin{aligned}
F_{max}(y) &= \mathbb{P}((x_1 < y) \cap (x_2 < y) \cap \ldots \cap (x_n < y)) \\
. &= \prod_i F(y) = F^n(y) = \frac{(y-a)^n}{(b-a)^n}
\end{aligned}
$$

# Maximum Likelihood Estimation, MLE



- Calculate $\langle max(x) \rangle$ from the CDF!
- Utilize the independence of data points!

$$
\begin{aligned}
F_{max}(y) &= \mathbb{P}((x_1 < y) \cap (x_2 < y) \cap \ldots \cap (x_n < y)) \\
. &= \prod_i F(y) = F^n(y) = \frac{(y-a)^n}{(b-a)^n}
\end{aligned}
$$

- Expectation value for max(x):

$$
\langle max(x) \rangle = \int_a^b dy \; yn \frac{(y-a)^{n-1}}{(b-a)^n} = \frac{n}{n+1}(b-a) + a
$$

# Maximum Likelihood Estimation, MLE

- Calculate $\langle max(x) \rangle$ from the CDF!
- Utilize the independence of data points!

$$
\begin{aligned}
F_{max}(y) &= \mathbb{P}((x_1 < y) \cap (x_2 < y) \cap \ldots \cap (x_n < y)) \\
. &= \prod_i F(y) = F^n(y) = \frac{(y-a)^n}{(b-a)^n}
\end{aligned}
$$

- Expectation value for max(x):

$$
\langle max(x) \rangle = \int_a^b dy\; yn \frac{(y-a)^{n-1}}{(b-a)^n} = \frac{n}{n+1}(b-a) + a
$$

- From this we have an estimate for $b$, assuming $a$ is known:

$$
b = \frac{n+1}{n}(max(x) - a) + a
$$

# Maximum Likelihood Estimation, MLE



- Calculate $\langle max(x) \rangle$ from the CDF! Utilize the independence of data points!

$$
\begin{aligned}
F_{max}(y) &= \mathbb{P}((x_1 < y) \cap (x_2 < y) \cap \ldots \cap (x_n < y)) \\
. &= \prod_i F(y) = F^n(y) = \frac{(y-a)^n}{(b-a)^n}
\end{aligned}
$$

- Expectation value for max(x):

$$
\langle max(x) \rangle = \int_a^b dy\ yn \frac{(y-a)^{n-1}}{(b-a)^n} = \frac{n}{n+1}(b-a) + a
$$

- From this we have an estimate for $b$, assuming $a$ is known:

$$
b = \frac{n+1}{n}(max(x) - a) + a
$$

- Because we do not know $a$ at the beginning, we use the MLE value for $a$. Then (using a similar formula for the minimum), the estimated $b$ will be used for $a$. Iterate this procedure until convergence!

# Bootstrap

- Calculate the 95% confidence interval for the median of experimental data! Apply bootstrap method with percentile calculation! (sheet: "boot")

# Bootstrap

- Calculate the 95% confidence interval for the median of experimental data! Apply bootstrap method with percentile calculation! (sheet: "boot")
- Create a sample with replacement, and calculate the median for the sample! Repeat 2000 times!

# Bootstrap

- Calculate the 95% confidence interval for the median of experimental data! Apply bootstrap method with percentile calculation! (sheet: "boot")
- Create a sample with replacement, and calculate the median for the sample! Repeat 2000 times!
- Create PDF for the median values from the bootstrapped sample,calculate the percentile values. Remind: confidence intervals are two sided by default.

# Multiple testing



- Problem:
  - We have 25 disease indicators for a child. We want to conduct an experiment, where we need a healthy subject, none of the indicators are allowed to show a disease. Our lab is able to test **each indicator** at $0.95$ confidence against the disease. We found some indicators, where the test for being healthy was rejected. Should we discard this subject, if our threshold is 95% confidence level for conducting the experiment? (sheet: "bonf")

# Multiple testing



- Problem:
  - We have 25 disease indicators for a child. We want to conduct an experiment, where we need a healthy subject, none of the indicators are allowed to show a disease. Our lab is able to test **each indicator** at $0.95$ confidence against the disease. We found some indicators, where the test for being healthy was rejected. Should we discard this subject, if our threshold is 95% confidence level for conducting the experiment? (sheet: "bonf")
- Solution:

# Multiple testing



- Problem:
  - We have 25 disease indicators for a child. We want to conduct an experiment, where we need a healthy subject, none of the indicators are allowed to show a disease. Our lab is able to test **each indicator** at $0.95$ confidence against the disease. We found some indicators, where the test for being healthy was rejected. Should we discard this subject, if our threshold is 95% confidence level for conducting the experiment? (sheet: "bonf")
- Solution:
  - Bonferroni method

# Multiple testing



- Problem:
  - We have 25 disease indicators for a child. We want to conduct an experiment, where we need a healthy subject, none of the indicators are allowed to show a disease. Our lab is able to test **each indicator** at $0.95$ confidence against the disease. We found some indicators, where the test for being healthy was rejected. Should we discard this subject, if our threshold is 95% confidence level for conducting the experiment? (sheet: "bonf")
- Solution:
  - Bonferroni method
  - Benjamini-Hochenberg method

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?
- We have $m$ separate hypothesis tests

$$H_0^i \leftrightarrow H_1^i \text{ with } p - \text{value } P_i$$

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?
- We have $m$ separate hypothesis tests

$$H_0^i \leftrightarrow H_1^i \text{ with } p - \text{value } P_i$$

- Bonferroni correction:

$$\text{reject } H_0^i \text{ if } P_i < \alpha/m$$

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?
- We have $m$ separate hypothesis tests

$$H_0^i \leftrightarrow H_1^i \text{ with } p - \text{value } P_i$$

- Bonferroni correction:

$$\text{reject } H_0^i \text{ if } P_i < \alpha/m$$

Proof: $\mathbb{P}$(any test $A_i$ falsely rejected)
$= \mathbb{P}(\cup_i^m A_i) \leq \sum_i^m \mathbb{P}(A_i) = \sum_i^m \alpha/m = \alpha$

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?
- We have $m$ separate hypothesis tests

$$H_0^i \leftrightarrow H_1^i \text{ with } p-\text{value } P_i$$

- Bonferroni correction:

$$\text{reject } H_0^i \text{ if } P_i < \alpha/m$$

Proof: $\mathbb{P}$(any test $A_i$ falsely rejected)
$= \mathbb{P}(\cup_i^m A_i) \leq \sum_i^m \mathbb{P}(A_i) = \sum_i^m \alpha/m = \alpha$

- This correction is very conservative.

# Multiple testing: Bonferroni correction



- How to set the confidence level for **each test** if we want to ensure the probability of false rejection of **any** null hypothesis is less than $\alpha$?
- We have $m$ separate hypothesis tests

$$H_0^i \leftrightarrow H_1^i \text{ with } p - \text{value } P_i$$

- Bonferroni correction:

$$\text{reject } H_0^i \text{ if } P_i < \alpha/m$$

Proof: $\mathbb{P}$(any test $A_i$ falsely rejected)
$= \mathbb{P}(\cup_i^m A_i) \leq \sum_i^m \mathbb{P}(A_i) = \sum_i^m \alpha/m = \alpha$

- This correction is very conservative.
- The power of this correction is quite low.

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  - * order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  - order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$
  - set temporary threshold $t_i = i \frac{1}{C_m} \frac{\alpha}{m}$ for all tests, where $C_m = \sum_{i=1}^{m} 1/i$

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  - * order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$
  - * set temporary threshold $t_i = i \frac{1}{C_m} \frac{\alpha}{m}$ for all tests, where $C_m = \sum_{i=1}^{m} 1/i$
  - * Starting from the largest $P_i$ and going down, find the first $P_j < t_j$

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  - * order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$
  - * set temporary threshold $t_i = i \frac{1}{C_m} \frac{\alpha}{m}$ for all tests, where $C_m = \sum_{i=1}^{m} 1/i$
  - * Starting from the largest $P_i$ and going down, find the first $P_j < t_j$
  - * reject all $H_0$ with lower p-value.

# Multiple testing: B-H correction

Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  * order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$
  * set temporary threshold $t_i = i \frac{1}{C_m} \frac{\alpha}{m}$ for all tests, where $C_m = \sum_{i=1}^{m} 1/i$
  * Starting from the largest $P_i$ and going down, find the first $P_j < t_j$
  * reject all $H_0$ with lower p-value.
- The B-H correction is more popular in case of testing a large numberof hypotheses.

# Multiple testing: B-H correction



## Benjamini-Hochberg correction

- Allow some more false rejection, but control the False Discovery Rate (FDR):

$$FDR = \left\langle \frac{\text{number of false rejections}}{\text{number of all rejections}} \right\rangle \leq \alpha$$

- The correction procedure:
  - * order the tests by increasing p-values: $P_i \leq P_j$ for $i < j$
  - * set temporary threshold $t_i = i \frac{1}{C_m} \frac{\alpha}{m}$ for all tests, where $C_m = \sum_{i=1}^{m} 1/i$
  - * Starting from the largest $P_i$ and going down, find the first $P_j < t_j$
  - * reject all $H_0$ with lower p-value.
- The B-H correction is more popular in case of testing a large numberof hypotheses.
- The correction factor for independent tests is smaller: $C_m \leq 1$

# Independence



Main concept:

- two by two table of outcomes

| $.$ | $a=0$ | $a=1$ | $.$ |
|---|---|---|---|
| $b=0$ | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| $b=1$ | $p_{10}$ | $p_{11}$ | $p_{1.}$ |
| $.$ | $p_{.0}$ | $p_{.1}$ | $p_{..}=1$ |

# Independence

Main concept:

- two by two table of outcomes

| . | $a=0$ | $a=1$ | . |
|---|---|---|---|
| $b=0$ | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| $b=1$ | $p_{10}$ | $p_{11}$ | $p_{1.}$ |
| . | $p_{.0}$ | $p_{.1}$ | $p_{..}=1$ |

- Odds ratio:

$$OR = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

# Independence



Main concept:

- two by two table of outcomes

| . | $a=0$ | $a=1$ | . |
|---|---|---|---|
| $b=0$ | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| $b=1$ | $p_{10}$ | $p_{11}$ | $p_{1.}$ |
| . | $p_{.0}$ | $p_{.1}$ | $p_{..}=1$ |

- Odds ratio:

$$OR = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

- a and b are independent iff:

$$OR = 1$$

or equivalently

$$p_{ij} = p_{i.}p_{.j}$$

# Independence

Main concept:

- two by two table of outcomes

| . | $a=0$ | $a=1$ | . |
|---|---|---|---|
| $b=0$ | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| $b=1$ | $p_{10}$ | $p_{11}$ | $p_{1.}$ |
| . | $p_{.0}$ | $p_{.1}$ | $p_{..}=1$ |

- Odds ratio:

$$OR = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

- a and b are independent iff:

$$OR = 1$$

or equivalently

$$p_{ij} = p_{i.}p_{.j}$$

- Main tests:

- $\chi^2$ sheet: "chi2"
- Fisher-exact, sheet: "fish"