Chapter 2: Basics from Probability Theory and Statistics

2.1 Probability Theory

Events, Probabilities, Random Variables, Distributions, Moments

Generating Functions, Deviation Bounds, Limit Theorems

Basics from Information Theory

2.2 Statistical Inference: Sampling and Estimation

Moment Estimation, Confidence Intervals

Parameter Estimation, Maximum Likelihood, EM Iteration

2.3 Statistical Inference: Hypothesis Testing and Regression

Statistical Tests, p-Values, Chi-Square Test

Linear and Logistic Regression

mostly following L. Wasserman, with additions from other sources

2.2 Statistical Inference:
Sampling and Estimation

A ***statistical model*** is a set of distributions (or regression functions), e.g., all unimodal, smooth distributions.
A ***parametric model*** is a set that is completely described by a finite number of parameters,
(e.g., the family of Normal distributions).

Statistical inference: given a sample X_1, ..., X_n how do we infer the distribution or its parameters within a given model.

For multivariate models with one specific „outcome (response)“ variable Y, this is called ***prediction*** or ***regression,***
for discrete outcome variable also ***classification***.
r(x) = E[Y | X=x] is called the ***regression function***.

Statistical Estimators

A ***point* estimator** for a parameter θ of a prob. distribution is a random variable X derived from a random sample $X_1, ..., X_n$.

Examples:

Sample mean: $$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample variance: $$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

An estimator T for parameter θ is ***unbiased***

if $E[T] = \theta$;

otherwise the estimator has bias $E[T] - \theta$.

An estimator on a sample of size n is ***consistent***

if $lim_{n\to\infty} P[|T - \theta| < \varepsilon] = 1 \;\; for\ each\ \varepsilon > 0$

Sample mean and sample variance
are unbiased, consistent estimators with minimal variance.

Estimator Error

Let $\hat{\theta}_n = T(\theta)$ be an estimator for parameter θ over sample $X_1, ..., X_n$.

The distribution of $\hat{\theta}_n$ is called the sampling distribution.

The ***standard error*** for $\hat{\theta}_n$ is: $se(\hat{\theta}) = \sqrt{Var[\hat{\theta}]}$

The ***mean squared error (MSE)*** for $\hat{\theta}_n$ is:

$$MSE(\hat{\theta}) = E[(\hat{\theta}_n - \theta)^2]$$
$$= bias^2(\hat{\theta}_n) + Var[\hat{\theta}_n]$$

If bias $\rightarrow 0$ and se $\rightarrow 0$ then the estimator is consistent.

The estimator $\hat{\theta}_n$ is ***asymptotically Normal*** if
$(\hat{\theta}_n - \theta)/se$ converges in distribution to standard Normal N(0,1)

Nonparametric Estimation

The ***empirical distribution function*** $\hat{F}_n$ is the cdf that puts prob. mass 1/n at each data point X_i: $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$

A ***statistical functional*** T(F) is any function of F,
e.g., mean, variance, skewness, median, quantiles, correlation

The ***plug-in estimator*** of $\theta = T(F)$ is: $\hat{\theta}_n = T(\hat{F}_n)$

Instead of the full empirical distribution, often compact data synopses may be used, such as ***histograms*** where $X_1, ..., X_n$ are grouped into m cells (buckets) $c_1, ..., c_m$ with bucket boundaries $lb(c_i)$ and $ub(c_i)$ s.t. $lb(c_1) = -\infty$, $ub(c_m) = \infty$, $ub(c_i) = lb(c_{i+1})$ for $1 \leq i < m$, and
$freq(c_i) = \hat{F}_n(x) = \frac{1}{n}\sum_{\nu=1}^{n} I(lb(c_i) \leq X_\nu < ub(c_i))$

Histograms provide a (discontinuous) ***density estimator***.

Parametric Inference: Method of Moments

Compute sample moments: $\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ for j-th moment α_j

Estimate parameter θ by method-of-moments estimator $\hat{\theta}_n$ s.t.

$$\alpha_1(F(\hat{\theta}_n)) = \hat{\alpha}_1$$

and $$\alpha_2(F(\hat{\theta}_n)) = \hat{\alpha}_2$$

and $$\alpha_3(F(\hat{\theta}_n)) = \hat{\alpha}_3$$

and **…** (for some number of moments)

Method-of-moments estimators are usually consistent and asympotically Normal, but may be biased

Parametric Inference: Maximum Likelihood Estimators (MLE)

Estimate parameter θ of a postulated distribution $f(\theta,x)$ such that the probability that the data of the sample are generated by this distribution is maximized.

$\rightarrow$ **Maximum likelihood estimation:**

Maximize $L(x_1,...,x_n, \theta) = P[x_1, ..., x_n \text{ originate from } f(\theta,x)]$

(often written as

$L(\theta \mid x_1,...,x_n) = f(x_1,...,x_n \mid \theta)$)

or maximize log L

if analytically untractable $\rightarrow$ use numerical iteration methods

MLE Properties

Maximum Likelihood Estimators are
consistent, asymptotically Normal, and
asymptotically optimal in the following sense:

Consider two estimators U and T which are asymptotically Normal.
Let u^2 and t^2 denote the variances of the two Normal distributions
to which U and T converge in probability.
The asymptotic relative efficiency of U to T is ARE(U,T) = t^2/u^2 .

Theorem: For an MLE $\hat{\theta}_n$ and any other estimator $\tilde{\theta}_n$
the following inequality holds:

$$ARE(\tilde{\theta}_n, \hat{\theta}_n) \leq 1$$

Simple Example for Maximum Likelihood Estimator

given:

- coin with Bernoulli distribution with unknown parameter p für head, 1-p for tail
- sample (data): k times head with n coin tosses

needed: maximum likelihood estimation of p

Let L(k, n, p) = P[sample is generated from distr. with param. p]

$$= \binom{n}{k} p^k \ (1-p)^{n-k}$$

Maximize log-likelihood function log L (k, n, p):

$$\log \quad \mathrm{L} = \log \binom{\mathrm{n}}{\mathrm{k}} \quad + \quad \mathrm{k} \quad \log \mathrm{p} \quad + \quad (\mathrm{n}-\mathrm{k}) \quad \log \quad (1-\mathrm{p})$$

$$\frac{\partial \log L}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \qquad \Rightarrow p = \frac{k}{n}$$

Advanced Example for Maximum Likelihood Estimator

given:

- Poisson distribution with parameter λ (expectation)
- sample (data): numbers $x_1, ..., x_n \in N_0$

needed: maximum likelihood estimation of λ

Let r be the largest among these numbers,
and let $f_0, ..., f_r$ be the absolute frequencies of numbers 0, ..., r.

$$L(x_1,...,x_n,\lambda) = \prod_{i=0}^{r} \left(e^{-\lambda} \frac{\lambda^i}{i!} \right)^{f_i}$$

$$\Rightarrow \frac{\partial \ln L}{\partial \lambda} = \sum_{i=0}^{r} f_i \left(\frac{i}{\lambda} - 1 \right) = 0 \qquad \Rightarrow \hat{\lambda} = \frac{\sum_{i=0}^{r} i\, f_i}{\sum_{i=0}^{r} f_i} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Sophisticated Example for Maximum Likelihood Estimator

given:

- discrete uniform distribution over $[1,\theta] \subseteq N_0$ and density $f(x) = 1/\theta$
- sample (data): numbers $x_1, ..., x_n \in N_0$

MLE for θ is $\max\{x_1, ..., x_n\}$ (see Wasserman p. 124)

MLE for Parameters of Normal Distributions

$$L(x_1,\ldots,x_n,\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\frac{\partial \ln(L)}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu) = 0$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

Bayesian Viewpoint of Parameter Estimation

- assume prior distribution $f(\theta)$ of parameter θ
- choose statistical model (generative model) $f(x \mid \theta)$ that reflects our beliefs about RV X
- given RVs $X_1, ..., X_n$ for observed data, the posterior distribution is $f(\theta \mid x_1, ..., x_n)$

for $X_1=x_1, ..., X_n=x_n$ the likelihood is

$$L(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} \frac{f(\theta \mid x_i) \cdot \sum_{\theta'} f(x_i \mid \theta') f(\theta')}{f(\theta)}$$

which implies

$$f(\theta \mid x_1,...,x_n) \sim L(x_1,...,x_n \mid \theta) \cdot f(\theta)$$

(posterior is proportional to likelihood times prior)

MAP estimator (maximum a posteriori):

compute θ that maximizes $f(\theta \mid x_1, \ldots, x_n)$ *given a prior for* θ

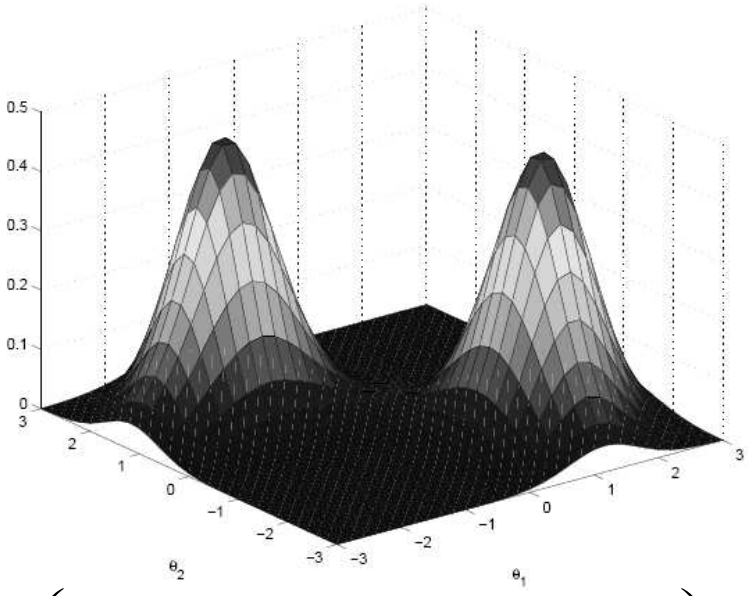
Analytically Non-tractable MLE for parameters of Multivariate Normal Mixture

consider samples from a mixture of multivariate Normal distributions with the density (e.g. height and weight of males and females):

$$f(\vec{x}, \pi_1, ..., \pi_k, \vec{\mu}_1, ..., \vec{\mu}_k, \Sigma_1, ..., \Sigma_k)$$

$$= \sum_{j=1}^{k} \pi_j \; n(\vec{x}, \vec{\mu}_j, \Sigma_j) = \sum_{j=1}^{k} \pi_j \; \frac{1}{\sqrt{(2\pi)^m \left|\Sigma_j\right|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}-\vec{\mu}_j)}$$

with expectation values $\vec{\mu}_j$
and invertible, positive definite, symmetric m×m covariance matrices Σ_j



$\rightarrow$ maximize log-likelihood function:

$$\log L(\vec{x}_1, ..., \vec{x}_n, \theta) := \log \prod_{i=1}^{n} P[\vec{x}_i \mid \theta] = \sum_{i=1}^{n} \left(\log \sum_{j=1}^{k} \pi_j \; n(\vec{x}_i, \vec{\mu}_j, \Sigma_j) \right)$$

Expectation-Maximization Method (EM)

Key idea:

when $L(\theta, X_1, ..., X_n)$ (where the X_i and θ are possibly multivariate)

is analytically intractable then

- introduce ***latent (hidden, invisible, missing) random variable(s) Z***

  such that

  - the ***joint distribution $J(X_1, ..., X_n, Z, \theta)$ of the „complete" data***

    is tractable (often with Z actually being $Z_1, ..., Z_n$)

- derive the incomplete-data likelihood $L(\theta, X_1, ..., X_n)$ by

  integrating (marginalization) J:

$$\hat{\theta} = arg\,max_{\theta} \sum_{z} J[\theta, X_1, ..., X_n, Z \mid Z = z] P[Z = z]$$

EM Procedure

Initialization: choose start estimate for $\theta^{(0)}$

Iterate (t=0, 1, …) until convergence:

E step (expectation):
estimate posterior probability of Z: $P[Z \mid X_1, \ldots, X_n, \theta^{(t)}]$
assuming θ were known and equal to previous estimate $\theta^{(t)}$,
and compute $E_{Z \mid X1, \ldots, Xn, \theta(t)}[\log J(X_1, \ldots, X_n, Z \mid \theta)]$
by integrating over values for Z

M step (maximization, MLE step):
Estimate $\theta^{(t+1)}$ by maximizing
$E_{Z \mid X1, \ldots, Xn, \theta(t)}[\log J(X_1, \ldots, X_n, Z \mid \theta)]$

convergence is guaranteed
(because the E step computes a lower bound of the true L function,
and the M step yields monotonically non-decreasing likelihood),
but may result in local maximum of log-likelihood function

EM Example for Multivariate Normal Mixture

Expectation step (E step):

$$h_{ij} := P[Z_{ij} = 1 \mid \vec{x}_i, \theta^{(t)}] = \frac{P[\vec{x}_i \mid n_j(\theta^{(t)})]}{\sum_{l=1}^{k} P[\vec{x}_i \mid n_l(\theta^{(t)})]}$$

$Z_{ij} = 1$ *if* i^{th} *data point was generated by* j^{th} *component, 0 otherwise*

Maximization step (M step):

$$\vec{\mu}_j := \frac{\sum_{i=1}^{n} h_{ij}\vec{x}_i}{\sum_{i=1}^{n} h_{ij}} \qquad \Sigma_j := \frac{\sum_{i=1}^{n} h_{ij}(\vec{x}_i - \vec{\mu}_j)(\vec{x}_i - \vec{\mu}_j)^T}{\sum_{i=1}^{n} h_{ij}}$$

$$\pi_j := \frac{\sum_{i=1}^{n} h_{ij}}{\sum_{j=1}^{k}\sum_{i=1}^{n} h_{ij}} = \frac{\sum_{i=1}^{n} h_{ij}}{n}$$

$\Big\} \; \theta^{(t+1)}$

Confidence Intervals

Estimator T for an interval for parameter θ such that

$$P[T - a \leq \theta \leq T + a] = 1 - \alpha$$

[T-a, T+a] is the **confidence interval** and 1-α is the **confidence level.**

For the distribution of random variable X a value
x_γ (0< γ <1) with $P[X \leq x_\gamma] \geq \gamma \wedge P[X \geq x_\gamma] \geq 1 - \gamma$
is called a **γ quantile**; the 0.5 quantile is called the **median**.
For the normal distribution N(0,1) the γ quantile is denoted Φ_γ **.**

Confidence Intervals for Expectations (1)

Let $x_1, ..., x_n$ be a sample from a distribution with unknown expectation μ and known variance σ^2.

For sufficiently large n the sample mean $\overline{X}$ is $N(\mu,\sigma^2/n)$ distributed and $\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}$ is N(0,1) distributed:

$$P[-z \le \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} \le z] = \Phi(z) - \Phi(-z) = \Phi(z) - (1-\Phi(z))$$

$$= P[\overline{X} - \frac{z\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{z\sigma}{\sqrt{n}}]$$

$$\Rightarrow P[\overline{X} - \frac{\Phi_{1-\alpha/2}\,\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{\Phi_{1-\alpha/2}\,\sigma}{\sqrt{n}}] = 1-\alpha$$

For required confidence interval $[\overline{X}-a, \overline{X}+a]$ or confidence level $1-\alpha$ set

$$z := \frac{a\sqrt{n}}{\sigma}$$

then look up $\Phi(z)$

or

$$z := (1-\frac{\alpha}{2}) \text{ quantile of } N(0,1)$$

then $a := \frac{z\sigma}{\sqrt{n}}$

Confidence Intervals for Expectations (2)

Let $x_1, ..., x_n$ be a sample from a distribution with unknown expectation μ and *unknown variance* σ^2 and sample variance S^2 .
For sufficiently large n the random variable

$$T := \frac{(\bar{X} - \mu)\sqrt{n}}{S}$$ has a t distribution (Student distribution) with n-1 degrees of freedom:

$$f_{T,n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{n\pi}\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

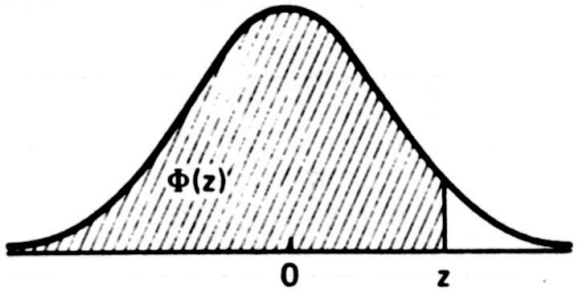
with the Gamma function: $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \quad \text{für } x > 0$

(with the properties $\Gamma(1) = 1$ *and* $\Gamma(x+1) = x\Gamma(x)$ *)*

$$\Rightarrow P[\bar{X} - \frac{t_{n-1,1-\alpha/2}\, S}{\sqrt{n}} \le \mu \le \bar{X} + \frac{t_{n-1,1-\alpha/2}\, S}{\sqrt{n}}] = 1-\alpha$$

Normal Distribution Table

The Normal Distribution Functions $\Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$

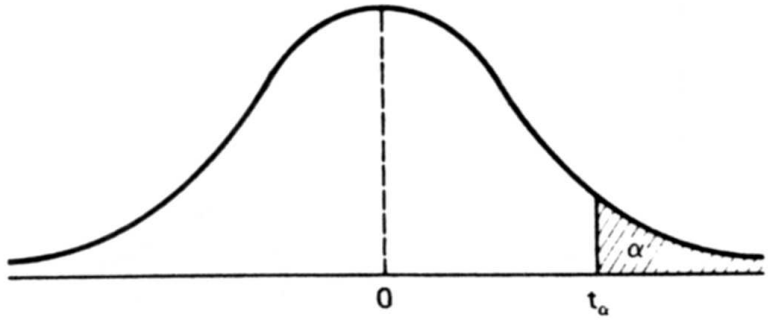


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158

Student‘s t Distribution Table

Table 5

Critical Values of the Student-*t* Distribution[a]



n \ α	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898

2.3 Statistical Inference:
Hypothesis Testing and Regression

Hypothesis testing:

- aims to falsify some hypothesis by lack of statistical evidence
- design of test RV and its (approximate / limit) distribution

Regression:

- aims to estimate joint distribution of input and output RVs based on some model and usually minimizing quadratic error

Statistical Hypothesis Testing

A hypothesis test determines a probability 1-α (***test level α, significance level***) that a sample $X_1, ..., X_n$ from some unknown probability distribution has a certain property.

Examples:

1) The sample originates from a normal distribution.
2) Under the assumption of a normal distribution the sample originates from a $N(\mu, \sigma^2)$ distribution.
3) Two random variables are independent.
4) Two random variables are identically distributed.
5) Parameter λ of a Poisson distribution from which the sample stems has value 5.
6) Parameter p of a Bernoulli distribution from which the sample stems has value 0.5.

General form:

null hypothesis H_0 vs. alternative hypothesis H_1

needs test variable X (derived from $X_1, ..., X_n, H_0, H_1$) and test region R with

$X \in R$ for rejecting H_0 and

$X \notin R$ for retaining H_0

	Retain H_0	Reject H_0
H_0 true	✓	type I error
H_1 true	type II error	✓

Hypotheses and p-Values

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.
A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called composite hypothesis.
A test of the form $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$ is called a two-sided test.
A test of the form $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$ or $H_0: \theta \geq \theta_0$ vs. $H_1: \theta < \theta_0$ is called a one-sided test.

Suppose that for every level $\alpha \in (0,1)$ there is a test with rejection region R_α. Then the ***p-value*** is the smallest level at which we can reject H_0: $p\text{-}value = inf\{\alpha \mid T(X_1,\ldots,X_n) \in R_\alpha\}$

small p-value means strong evidence against H_0

Hypothesis Testing Example

Null hypothesis for n coin tosses: coin is fair or has head probability $p = p_0$; ***alternative hypothesis***: $p \neq p_0$

Test variable: X, the #heads, is

$N(pn, p(1-p)n^2)$ distributed (by the Central Limit Theorem),

thus $$Z := \frac{X/n - p}{\sqrt{p(1-p)}}$$ is N(0, 1) distributed

Rejection of null hypothesis at test level α (e.g. 0.05) if

$$Z > \Phi_{1-\alpha/2} \vee Z < \Phi_{\alpha/2}$$

Wald Test

for testing $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$ use the test variable $W = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

with sample estimate $\hat{\theta}$ and standard error $se(\hat{\theta}) = \sqrt{Var[\hat{\theta}]}$

W converges in probability to N(0,1)

$\rightarrow$ reject H_0 at level α when $|W| > \Phi_{\alpha/2}$

Chi-Square Distribution

Let $X_1, ..., X_n$ be independent, N(0,1) distributed random variables.
Then the random variable $\chi_n^2 := X_1^2 + ... + X_n^2$
is chi-square distributed with n degrees of freedom:

$$f_{\chi_n^2}(x) = \frac{x^{\frac{n-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \text{ for } x > 0, \ 0 \text{ otherwise}$$

Let n be a natural number, let X be N(0,1) distributed and
Y χ^2 distributed with n degrees of freedom.
Then the random variable $T_n := \sqrt{n} \frac{X}{\sqrt{Y}}$
is t distributed with n degrees of freedom.

Chi-Square Goodness-of-Fit-Test

Given:

n sample values $X_1, ..., X_n$ of random variable X

with relative frequencies $H_1, ..., H_k$ for k value classes v_i

(e.g. value intervals) of random variable X

Null hypothesis:

the values X_i are f distributed (e.g. uniformly distributed),

where f has expectation μ and variance σ^2

Approach: $Y_k := \sum_{i=1}^{k} (H_i - E(v_i)) \sqrt{n} / \sigma$ and $Z_k := \sum_{i=1}^{k} \frac{(H_i - E(v_i)^2}{E(v_i)}$

with $E(v_i) := n\ P[X$ is in class v_i according to f $]$

are both approximately $\chi 2$ distributed with k-1 degrees of freedom

Rejection of null hypothesis at test level α (e.g. 0.05) if $Z_k > \chi^2_{k-1,1-\alpha}$

Chi-Square Independence Test

Given:

n samples of two random variables X, Y or, equivalently,
a twodimensional random variable
with (absolute) frequencies H_{11}, ..., H_{rc} for r*c value classes,
where X has r and Y has c distinct classes.
(This is called a ***contingency table***.)

Null hypothesis:

X und Y are independent; then the
expectations for the relative frequencies of the value classes would be

$$E_{ij} = \frac{R_i C_j}{n} \quad \text{with } R_i := \sum_{j=1}^{c} H_{ij} \quad \text{and} \quad C_j := \sum_{i=1}^{r} H_{ij}$$

Approach: $Z := \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(H_{ij} - E_{ij})^2}{E_{ij}}$ is approximately χ^2 distributed with (r-1)(c-1) degrees of freedom

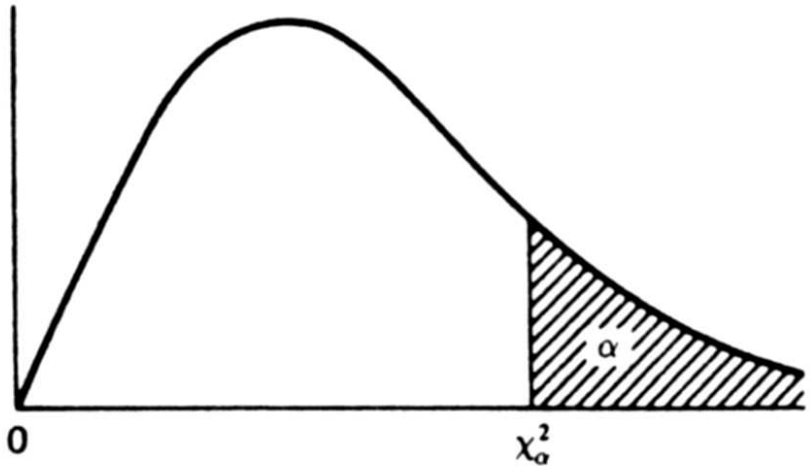
Rejection of null hypothesis at test level α (e.g. 0.05) if

$$Z > \chi^2_{(r-1)(c-1),\,1-\alpha}$$

Chi-Square Distribution Table

Table 4

Critical Values of the Chi-Square Distribution[a]



n^b \ α	0.995	0.990	0.975	0.950	0.05	0.025	0.010	0.005
1	0.0^4393^c	0.0^3157^c	0.0^3982^c	0.0^2393^c	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2103	10.597
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	11.071	12.833	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1674	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	15.507	17.535	20.090	21.955
9	1.7350	2.0879	2.7004	3.3251	16.920	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8158	4.5748	19.675	21.920	24.725	26.757
12	3.0738	3.5706	4.4038	5.2260	21.026	23.337	26.217	28.300
13	3.5650	4.1069	5.0087	5.8919	22.362	24.736	27.688	29.819

Chi-Square Distribution Table

1.1.2.10. Obere 100α-prozentige Werte χ^2_α der χ^2-Verteilung (s. 5.2.3.)

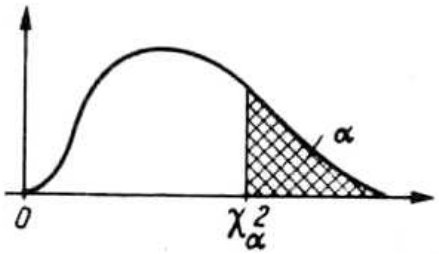


Abb. 1.4

Anzahl der Freiheitsgrade m	Wahrscheinlichkeit $p = \alpha$															
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07	1,64	2,7	3,8	5,4	6,6	7,9	9,5	10,83
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22	4,6	6,0	7,8	9,2	10,6	12,4	13,8
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,67	4,64	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,9	6,0	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3	9,2	11,1	13,4	15,1	16,8	18,9	20,5
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6	10,6	12,6	15,0	16,8	18,5	20,7	22,5
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,4	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0	13,4	15,5	18,2	20,1	22,0	24,3	26,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8	18,5	21,0	24,1	26,2	28,3	30,9	32,9
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2	21,1	23,7	26,9	29,1	31,3	34,0	36,1
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6	32,8	35,6	37,7
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0	34,3	37,1	39,3
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4	35,7	38,6	40,8
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8	37,2	40,1	42,3
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9	27,2	30,1	33,7	36,2	38,6	41,6	43,8
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6	40,0	43,0	45,3
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9	41,4	44,5	46,8
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3	42,8	45,9	48,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6	44,2	47,3	49,7
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0	45,6	48,7	51,2
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,6	44,3	46,9	50,1	52,6
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6	48,3	51,6	54,1
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	49,6	52,9	55,5
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	51,0	54,4	56,9

Linear Regression
(often used for parameter fitting of models)

Estimate r(x) = E[Y | X_1=x_1 ∧ ... ∧X_m=x_m] using a linear model

$$Y = r(x) + \varepsilon = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \varepsilon$$ with error ε with E[ε]=0

given n sample points ($x_1^{(i)}$, ..., $x_m^{(i)}$, $y^{(i)}$), i=1..n, the least-squares estimator (LSE) minimizes the quadratic error:

$$\sum_{i=1..n} \left(\left(\sum_{k=0..m} \beta_k x_k^{(i)} \right) - y^{(i)} \right)^2 =: E(\beta_0, ..., \beta_m)$$ (with $x_o^{(i)}$=1)

Solve linear equation system: $\frac{\partial E}{\partial \beta_k} = 0$ for k=0, ..., m

equivalent to MLE $\vec{\beta} = (X^T X)^{-1} X^T Y$

with Y = ($y^{(1)}$ ... $y^{(n)}$)^T and $X = \begin{pmatrix} 1\ x_1^{(1)}\ x_2^{(1)}\ \ldots\ x_m^{(1)} \\ 1\ x_1^{(2)}\ x_2^{(2)}\ \ldots\ x_m^{(2)} \\ \ldots \\ 1\ x_1^{(n)}\ x_2^{(n)}\ \ldots\ x_m^{(n)} \end{pmatrix}$

Logistic Regression

Estimate r(x) = E[Y | X=x] using a logistic model

$$Y = r(x) + \varepsilon = \frac{e^{\beta_0 + \sum_{i=1}^{m} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{m} \beta_i x_i}} + \varepsilon$$

with error ε with E[ε]=0

$\rightarrow$ solution for MLE for β_i values based on numerical methods

Additional Literature for Chapter 2

- Manning / Schütze: Chapters 2 und 6
- Duda / Hart / Stork: Appendix A
- R. Nelson: Probability, Stochastic Processes, and Queueing Theory, Springer, 1995
- M. Mitzenmacher, E. Upfal: Probability and Computing, Cambridge University Press, 2005
- M. Greiner, G. Tinhofer: Stochastik für Studienanfänger der Informatik, Carl Hanser Verlag, 1996
- G. Hübner: Stochastik, Vieweg, 1996
- Sean Borman: The Expectation Maximization Algorithm: A Short Tutorial, http://www.seanborman.com/publications/EM_algorithm.pdf
- Jason Rennie: A Short Tutorial on Using Expectation-Maximization with Mixture Models, http://people.csail.mit.edu/jrennie/writing/mixtureEM.pdf