II.2 Statistical Inference: Sampling and Estimation

A **statistical model M** is a set of distributions (or regression functions), e.g., all uni-modal, smooth distributions.

M is called a **parametric model** if it can be completely described by a finite number of parameters, e.g., the family of Normal distributions for a finite number of parameters μ, σ:

$$M = \left\{ f_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad for\ \mu \in R, \sigma > 0 \right\}$$

Statistical Inference

Given a parametric model M and a sample $X_1,\ldots,X_n$, how do we infer (learn) the parameters of M?

For multivariate models with observed variable X and „outcome (response)“ variable Y, this is called **prediction** or **regression**, for a discrete outcome variable this is also called **classification**.

$r(x) = E[Y \mid X=x]$ is called the **regression function**.

Idea of Sampling

Distribution X
(e.g., a population, objects of interest)

Statistical Inference
What can we say about X based on $X_1,\ldots,X_n$?

Samples $X_1,\ldots,X_n$ drawn from X
(e.g., people, objects)

Distrib. Param.		Sample Param.
μ_X	mean	$\overline{X}$
σ_X^2	variance	S_X^2
N	size	n

Example:

Suppose we want to estimate the average salary of employees in German companies.

→ Sample 1: Suppose we look at n=200 top-paid CEOs of major banks.

→ Sample 2: Suppose we look at n=100 employees across all kinds of companies.

Basic Types of Statistical Inference

Given a set of **iid. samples** $X_1,\ldots,X_n \sim X$
of an unknown distribution X.

e.g.: n single-coin-toss experiments $X_1,\ldots,X_n \sim X$: Bernoulli(p)

- **Parameter Estimation**

  e.g.: - what is the parameter p of X: Bernoulli(p) ?
  - what is E[X], the cdf F_X of X, the pdf f_X of X, etc.?

- **Confidence Intervals**

  e.g.: give me all values C=(a,b) such that $P(p \in C) \geq 0.95$
  where a and b are derived from samples $X_1,\ldots,X_n$

- **Hypothesis Testing**

  e.g.: $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

Statistical Estimators

A **point estimator** for a parameter θ of a prob. distribution X is a random variable $\hat{\theta}_n$ derived from an iid. sample $X_1,\ldots,X_n$.

<u>Examples:</u>

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

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

An estimator $\hat{\theta}_n$ for parameter θ is **unbiased**

if $E[\hat{\theta}_n] = \theta$;

otherwise the estimator has **bias** $E[\hat{\theta}_n] - \theta$.

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

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

Sample mean and sample variance
are unbiased and consistent estimators of μ_X and σ_X^2.

Estimator Error

Let $\hat{\theta}_n$ be an estimator for parameter θ over iid. samples $X_1, \ldots, X_n$.

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

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

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

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

<u>Theorem:</u> 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).

Types of Estimation

- **Nonparametric Estimation**

  No assumptions about model M nor the parameters θ of the underlying distribution X

  → "Plug-in estimators" (e.g. histograms) to approximate X

- **Parametric Estimation (Inference)**

  Requires assumptions about model M and the parameters θ of the underlying distribution X

  Analytical or numerical methods for estimating θ

  → Method-of-Moments estimator

  → Maximum Likelihood estimator and Expectation Maximization (EM)

Nonparametric Estimation

The **empirical distribution function** $\hat{F}_n$ is the cdf that puts probability mass *1/n* at each data point X_i:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x)$$

$$with \;\; I(X_i \le x) = \begin{cases} 1 \;\; if \;\; X_i \le x \\ 0 \;\; if \;\; X_i > x \end{cases}$$

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

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

→ Simply use $\hat{F}_n$ instead of F to calculate the statistics T of interest.

Histograms as Density Estimators

Instead of the full empirical distribution, often compact data synopses may be used, such as **histograms** where $X_1, \ldots, X_n$ are grouped into m cells (buckets) $c_1, \ldots, c_m$ with bucket boundaries $lb(c_i)$ and $ub(c_i)$ s.t.

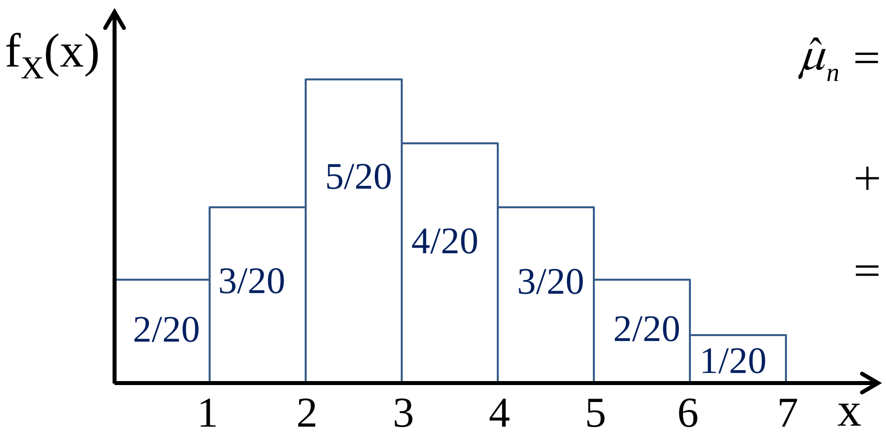
$lb(c_1) = -\infty,\ ub(c_m) = \infty,\ ub(c_i) = lb(c_{i+1})$ for $1 \le i < m$, and

$$freq_f(c_i) = \hat{f}_n(x) = \frac{1}{n}\sum_{v=1}^{n} I(lb(c_i) < X_v \le ub(c_i))$$

$$freq_F(c_i) = \hat{F}_n(x) = \frac{1}{n}\sum_{v=1}^{n} I(X_v \le ub(c_i))$$

Example:

$X_1 = 1$
$X_2 = 1$
$X_3 = 2$
$X_4 = 2$
$X_5 = 2$
$X_6 = 3$
…
$X_{20} = 7$



$$\hat{\mu}_n = 1\cdot\frac{2}{20} + 2\cdot\frac{3}{20} + 3\cdot\frac{5}{20} + 4\cdot\frac{4}{20} + 5\cdot\frac{3}{20} + 6\cdot\frac{2}{20} + 7\cdot\frac{1}{20} = 3.65$$

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

Parametric Inference (1): Method of Moments

Suppose parameter $\theta = (\theta_1,\ldots,\theta_k)$ has k components.

Compute **j-th moment:** $\alpha_j = \alpha_j(\theta) = E_\theta[X^j] = \int_{-\infty}^{+\infty} x^j \, f_X(x)\, dx$

j-th sample moment: $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ for $1 \le j \le k$

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

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

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

… …

and $\alpha_k(\hat{\theta}_n) = \hat{\alpha}_k$ (for the first k moments)

→ Solve equation system with k equations and k unknowns.

Method-of-moments estimators are usually **consistent** and **asymptotically Normal**, but may be **biased**.

Parametric Inference (2): Maximum Likelihood Estimators (MLE)

Let $X_1,\ldots,X_n$ be iid. with pdf $f(x;\theta)$.

Estimate parameter θ of a postulated distribution $f(x;\theta)$ such that the likelihood that the sample values $x_1,\ldots,x_n$ are generated by this distribution is maximized.

→ **Maximum likelihood estimation:**

Maximize $\mathbf{L(x_1,\ldots,x_n;\theta) \approx P[x_1,\ldots,x_n \text{ originate from } f(x;\theta)]}$

Usually formulated as

$\mathbf{L_n(\theta) = \prod_i f(X_i;\theta)}$

Or (alternatively)

→ **Maximize** $\mathbf{l_n(\theta) = \log L_n(\theta)}$

The value $\hat{\theta}_n$ that maximizes $L_n(\theta)$ is the **MLE** of θ.

If analytically untractable → use numerical iteration methods

Simple Example for Maximum Likelihood Estimator

<u>Given:</u>

- Coin toss experiment (Bernoulli distribution) with unknown parameter p for seeing heads, 1-p for tails
- Sample (data): h times head with n coin tosses

<u>Want:</u> Maximum likelihood estimation of p

Let $L(h, n, p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^h(1-p)^{n-h}$

$with\ h = \sum_i X_i$

<u>Maximize log-likelihood function:</u>

$$log\ L\ (h, n, p) = h\log(p) + (n-h)\log(1-p)$$

$$\frac{\partial \ln L}{\partial p} = \frac{h}{p} - \frac{n-h}{1-p} = 0 \quad \Rightarrow p = \frac{h}{n}$$

MLE for Parameters of Normal Distributions

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

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

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

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

MLE Properties

Maximum Likelihood estimators are
consistent, **asymptotically Normal**, and
asymptotically optimal (i.e., **efficient**) 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:

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

That is, among all estimators MLE has the smallest variance.

Bayesian Viewpoint of Parameter Estimation

- Assume **prior distribution** $g(\theta)$ of parameter θ
- Choose statistical model (**generative model**) $f(x \mid \theta)$ that reflects our beliefs about RV X
- Given RVs $X_1,\ldots,X_n$ for the observed data, the **posterior distribution** is $h(\theta \mid x_1,\ldots,x_n)$

For $X_1 = x_1, \ldots, X_n = x_n$ the likelihood is

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

which implies

$$h(\theta \mid x_1 \ldots x_n) \sim L(x_1 \ldots x_n, \theta) \cdot g(\theta)$$

(posterior is proportional to likelihood times prior)

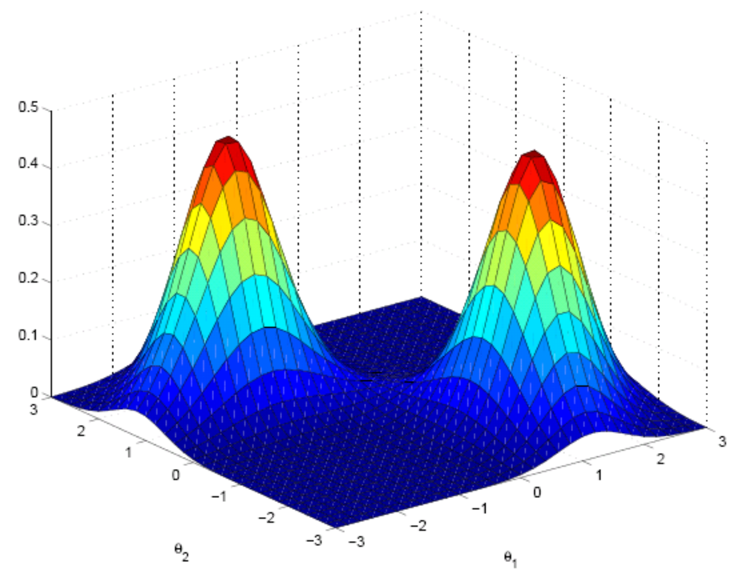
MAP estimator (maximum a posteriori):
Compute θ that <u>maximizes</u> $h(\theta \mid x_1, \ldots, x_n)$ ***given a prior*** *for* θ.

Analytically Non-tractable MLE for parameters of Multivariate Normal Mixture

Consider samples from a *k*-mixture of *m*-dimensional Normal distributions with the density (e.g. height and weight of males and females):

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

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



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

$\rightarrow$ Maximize log-likelihood function:

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

Expectation-Maximization Method (EM)

Key idea:

When $L(X_1,\ldots,X_n, \theta)$ (where the X_i and θ are possibly multivariate) is analytically intractable then

- introduce **latent** (i.e., hidden, invisible, missing) **random variable(s) Z** such that
  - the **joint distribution $J(X_1,\ldots,X_n, Z, \theta)$ of the "complete" data** is tractable (often with Z actually being multivariate: $Z_1,\ldots,Z_m$)
- iterately derive the **expected complete-data likelihood** by integrating J and find best θ:

$$\hat{\theta} = \arg\max_\theta \underbrace{\sum_z J[X_1 \ldots X_n, \theta \mid Z = z] P[Z = z]}_{E_{Z|X,\theta}[J(X_1,\ldots,X_n,\, Z,\, \theta)]}$$

EM Procedure

Initialization: choose start estimate for $\theta^{(0)}$
(e.g., using Method-of-Moments estimator)

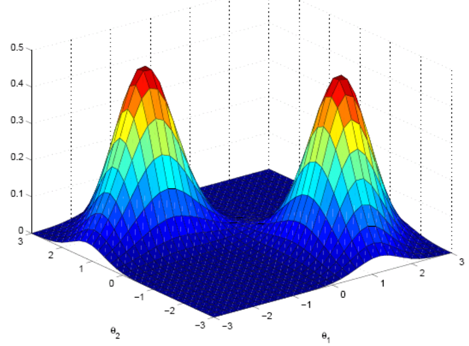
Iterate (t=0, 1, …) until convergence:

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 $E_{Z|X,\theta(t)}[\log J(X_1,\ldots,X_n, Z, \theta^{(t)})]$
by integrating over values for Z

M step (maximization, MLE step):
Estimate $\theta^{(t+1)}$ by maximizing
$\theta^{(t+1)} = \arg\max_\theta E_{Z|X,\theta}[\log J(X_1,\ldots,X_n, Z, \theta)]$

→ 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 X_i was generated by j^{th} component, 0 otherwise

Maximization step (M step):

$$E_{Z|X,\theta}[\log J(X_1, \ldots, X_n, Z, \theta)] = \sum_i \log \sum_j n_j(\vec{x}_i, \theta \mid Z_{ij} = 1) P[Z_{ij} = 1]$$

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

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

See L. Wasserman, p.121 ff.
for k=2, m=1

Confidence Intervals

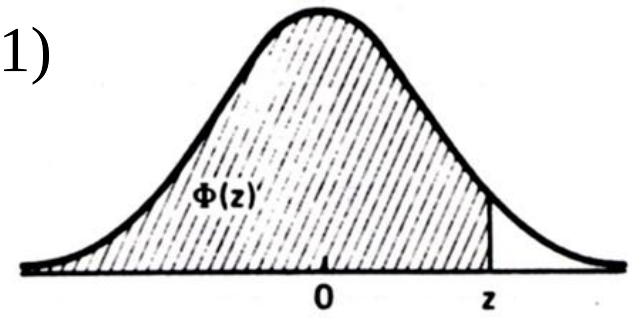
Estimator T for an interval for parameter θ such that

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

[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 $P[X \le x_\gamma] \ge \gamma \wedge P[X \ge x_\gamma] \ge 1 - \gamma$
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 .

→ For a given *a* or *α*, find a value z of N(0,1)
that denotes the [T-a, T+a] conf. interval
or a corresponding γ-quantile for 1–α .



Confidence Intervals for Expectations (1)

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

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

$$P[-z \le \frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \le z] = \Phi(z) - \Phi(-z) = \Phi(z) - (1-\Phi(z)) = 2\Phi(z) - 1$$

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

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

For **confidence interval** $[\bar{X}-a, \bar{X}+a]$ **or** **confidence level** 1-α set

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

then look up $\Phi(z)$ to find $1-\alpha$

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

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

Confidence Intervals for Expectations (2)

Let $X_1, \ldots, X_n$ be an iid. sample from a distribution X with *unknown expectation* μ and *unknown variance* σ^2 and *known 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**: $$\Gamma(x) = \int_0^\infty e^{-t}\, t^{x-1}\, dt \quad for\ x > 0$$

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

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

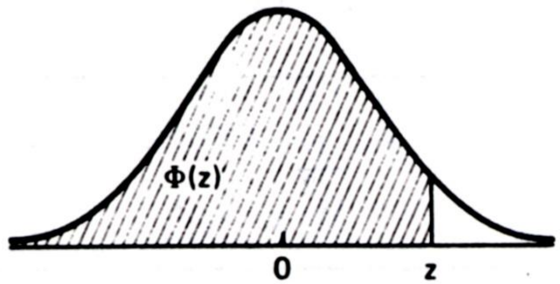
Summary of Section II.2

- **Quality measures** for statistical estimators
- **Nonparametric** vs. **parametric** estimation
- **Histograms** as generic (nonparametric) plug-in estimators
- **Method-of-Moments** estimator good initial guess but may be biased
- **Maximum-Likelihood estimator** & **Expectation Maximization**
- **Confidence intervals** for parameters

Normal Distribution Table

Table 3

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

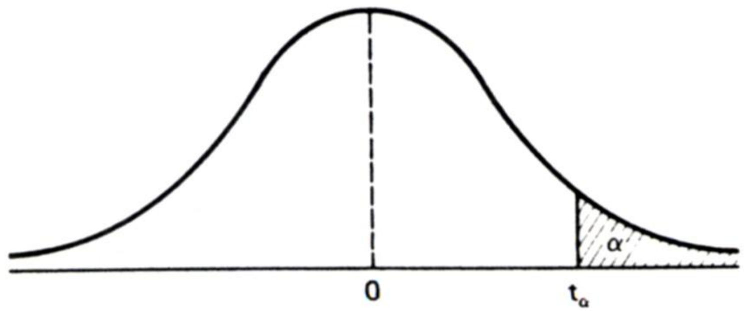


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.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
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

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
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845