Statistical Inference

Statistics is the method by which data are analyzed in whose generation chance is involved in some way.

Two main topics of statistics:

1. Estimation
2. Hypothesis testing

Statistical model:

- observed data $x = (x_1, \ldots, x_n)$
- x is a realization of the random variable $X = (X_1, \ldots, X_n)$
- $\mathcal{X}$ is the *sample space*, i.e., the set of possible observed data
- $\psi (\in \mathbb{R}^p)$ is a *parameter* which determines the distribution of X
- $\Psi (\subset \mathbb{R}^p)$ is the *parameter space*, i.e., the set of possible parameters ψ

Statistical Inference: Estimation

An *estimator* $\widehat{\psi}$ of the parameter ψ is some function of the random variable $(X_1, \ldots, X_n)$, i.e.,

$$\widehat{\psi} : \mathcal{X} \to \Psi$$

For observed data x, $\quad \widehat{\psi}(x)$ is called the *estimate* of Ψ.

Example (Coin tossing):

Suppose a coin is flipped n times. Let $X_i = 1$ (or 0), if trial i results in a head (or tail). Then, $X \in \{0, 1\}^n =: \mathcal{X}$. Let p denote the probability that trial i results in a head. Under the assumption that the trials are independent, p determines the distribution of $\mathcal{X}$, i.e., p is the parameter and $\Psi = [0, 1]$. Therefore, any function $\widehat{p} : \{0, 1\}^n \to [0, 1]$ is an estimator of p.

How to find a “good” estimator?

Likelihood $L(\psi \mid x)$

For given $x \in \mathcal{X}$, the likelihood function $L : \Psi \to \mathbb{R}$ is defined by

$L(\psi \mid x) = P_\psi(X = x)$ in case that the distribution of X is discrete or by

$L(\psi \mid x) = f_\psi(x)$ in case that the distribution of X is continuous with

probability density f_ψ.

If $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.), it follows that

$$L(\psi \mid x) = \begin{cases} \prod_{i=1}^n P_\psi(X_i = x_i) & \text{if } X_i \text{ are discrete r.v.'s} \\ \prod_{i=1}^n f_\psi(x_i) & \text{if } X_i \text{ are continuous r.v.'s} \end{cases}$$

Example (Coin tossing): Since $P_p(X_i = 1) = p$, it follows that

$$L(p \mid x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^r(1-p)^{n-r}$$

with $r := \sum_{i=1}^n x_i$.

Maximum likelihood estimator (MLE)

The maximum likelihood estimator is defined as

$$x \to \widehat{\psi}(x) := \arg\max_{\psi \in \Psi} L(\psi \mid x),$$

i.e., the estimate of ψ on the basis of observation x is the value of ψ which maximizes the likelihood.

Remark: It is often more convenient to maximize $\ln L(\psi \mid x)$ instead of $L(\psi \mid x)$. Since ln is a strict monotone function, the result is identical.

Example (Coin tossing):

$$\frac{d \ln L(p \mid x)}{dp} = \frac{r}{p} - \frac{n-r}{1-p} = 0 \quad \Leftrightarrow \quad r \cdot (1-p) - (n-r) \cdot p = 0$$
$$\Leftrightarrow \quad p = \frac{r}{n}$$

$\Rightarrow$ The MLE of p is the observed relative frequency of heads.

MLE: Multinomial distribution

$X_1, \ldots, X_n$ i.i.d. with $P(X_i = z_j) = p_j$ for $1 \leq j \leq k$ and $\sum_{j=1}^{k} p_j = 1$.

Then, $\mathcal{X} = \{z_1, \ldots, z_k\}^n$, $\psi = (p_1, \ldots, p_{k-1})$ and

$\Psi = \{(p_1, \ldots, p_{k-1}) : 0 \leq p_j \leq 1, \sum_{j=1}^{k-1} p_j \leq 1\}$.

The likelihood function $L : \Psi \to \mathbb{R}$ is given by

$$L(p_1, \ldots, p_{k-1} \mid x) = \left(\prod_{j=1}^{k-1} p_j^{r_j} \right) \cdot \underbrace{\left(1 - \sum_{j=1}^{k-1} p_j \right)}_{= p_k}^{r_k}$$

with $r_j := \sum_{i=1}^{n} 1_{(X_i = z_j)}$. Therefore,

$$\frac{d \ln L(p_1, \ldots, p_{k-1} \mid x)}{dp_s} = \frac{r_s}{p_s} - \frac{r_k}{1 - \sum_{j=1}^{k-1} p_j} = 0 \text{ for } 1 \leq s \leq k-1$$

$$\Leftrightarrow p_s = \frac{r_s}{n} \text{ for } 1 \leq s \leq k-1$$

MLE: Normal distribution

$X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$-distributed, i.e., $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is the density of the distribution of X_i. Then, $\mathcal{X} = \mathbb{R}^n$, $\psi = (\mu, \sigma^2)$, and $\Psi = \mathbb{R} \times \mathbb{R}^+$. The likelihood function $L : \Psi \to \mathbb{R}$ is given by $L(\mu, \sigma^2 \mid x) = \prod_{i=1}^n f(x_i)$.

$\Rightarrow$

$$
\begin{aligned}
\ln L(\mu, \sigma^2 \mid x) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}, \\
\frac{d \ln L(\mu, \sigma^2 \mid x)}{d\mu} &= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}, \\
\frac{d \ln L(\mu, \sigma^2 \mid x)}{d\sigma^2} &= -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^4}
\end{aligned}
$$

$\Rightarrow \quad (\widehat{\mu}, \widehat{\sigma}^2) := \left(\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n}\sum_{i=1}^n (x_i - \widehat{\mu})^2\right)$ is MLE.

Bias and mean square error (MSE)

The *bias* of an estimator $\widehat{\psi}$ is defined as $E_\psi \widehat{\psi} - \psi$.

An estimator $\widehat{\psi}$ is called *unbiased* in case that $E_\psi \widehat{\psi} - \psi = 0$ for all $\psi \in \Psi$.

Example (Normal distribution):

It can be shown that $E_{(\mu,\sigma^2)} \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu})^2 = \frac{n-1}{n}\sigma^2$. Therefore, the MLE of σ^2 is biased, whereas the estimator $\tilde{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \widehat{\mu})^2$ is unbiased.

The *mean square error* (MSE) of an estimator is defined as $MSE(\widehat{\psi}) := E_\psi(\widehat{\psi} - \psi)^2$. Since

$$E_\psi(\widehat{\psi} - \psi)^2 = E_\psi(\widehat{\psi} - E_\psi\widehat{\psi} + E_\psi\widehat{\psi} - \psi)^2 = \mathrm{Var}_\psi(\widehat{\psi}) + \left(E_\psi\widehat{\psi} - \psi\right)^2,$$

the mean square error of an estimator is the sum of its variance and its squared bias.

Computing maximum likelihood estimates

Example:

n unrelated individuals have been genotyped for two diallelic loci $\{A, a\}$ and $\{B, b\}$. Let n_{ij} denote the observed number of individuals possessing i alleles A and j alleles B $(0 \leq i, j \leq 2)$:

	bb	bB	BB
aa	n_{00} (ab/ab)	n_{01} (ab/aB)	n_{02} (aB/aB)
aA	n_{10} (ab/Ab)	n_{11} (ab/AB or aB/Ab)	n_{12} (aB/AB)
AA	n_{20} (Ab/Ab)	n_{21} (Ab/AB)	n_{22} (AB/AB)

Goal: Estimation of haplotype frequencies $p_{00} := P(ab)$, $p_{01} := P(aB)$, $p_{10} := P(Ab)$, and $p_{11} := P(AB)$ by the maximum likelihood method.

Computing maximum likelihood estimates

Example (continued):

$$
\begin{aligned}
\ln L(p_{00}, p_{01}, p_{10}, p_{11} \mid (n_{00}, \ldots, n_{11}, \ldots, n_{22})) \\
= \quad & (2n_{00} + n_{01} + n_{10}) \ln p_{00} + (n_{01} + 2n_{02} + n_{12}) \ln p_{01} \\
& + (n_{10} + 2n_{20} + n_{21}) \ln p_{10} + (n_{12} + n_{21} + 2n_{22}) \ln p_{11} \\
& + n_{11} \ln(p_{00} \cdot p_{11} + p_{01} \cdot p_{10}) + C
\end{aligned}
$$

Let $\tilde{n}_{11}$ denote the (unobserved) number of individuals with two-locus genotype ab/AB. In case that the “complete data” $(n_{00}, \ldots, \tilde{n}_{11}, n_{11} - \tilde{n}_{11}, \ldots, n_{22})$ were available, determination of MLEs would be straightforward (c.f. multinomial distribution): $\widehat{p}_{00} = \frac{2n_{00}+n_{01}+n_{10}+\tilde{n}_{11}}{2n}$ etc. On the other hand, if haplotype frequencies p_{ij} were known, the expected value of $\tilde{n}_{11}$, given n_{11} and p_{ij}, could easily be calculated: $E\tilde{n}_{11} = \frac{p_{00} \cdot p_{11}}{p_{00} \cdot p_{11} + p_{01} \cdot p_{10}} \cdot n_{11}$.

Computing MLEs: EM algorithm

Expectation-maximization (EM) algorithm:

0. Start with arbitrary values $p_{00}^{(0)}, \ldots, p_{11}^{(0)}$

For $r = 1, 2, \ldots$, repeat the following two steps

1. Expectation step: Calculate

$$E\tilde{n}_{11}^{(r)} = \frac{p_{00}^{(r-1)} \cdot p_{11}^{(r-1)}}{p_{00}^{(r-1)} \cdot p_{11}^{(r-1)} + p_{01}^{(r-1)} \cdot p_{10}^{(r-1)}} \cdot n_{11}$$

2. Maximization step: Calculate

$$p_{00}^{(r)} = \frac{2n_{00}+n_{01}+n_{10}+\tilde{n}_{11}^{(r)}}{2n}, \quad p_{01}^{(r)} = \frac{n_{01}+2n_{02}+n_{12}+n_{11}-\tilde{n}_{11}^{(r)}}{2n},$$

$$p_{10}^{(r)} = \frac{n_{10}+2n_{20}+n_{21}+n_{11}-\tilde{n}_{11}^{(r)}}{2n}, \quad p_{11}^{(r)} = \frac{n_{12}+n_{21}+2n_{22}+\tilde{n}_{11}^{(r)}}{2n}$$

until convergence occurs (e.g. $\max_{i,j} \mid p_{ij}^{(r)} - p_{ij}^{(r-1)} \mid \leq \varepsilon$).

EM algorithm: Pros and Cons

Pro:

- Easy to implement
- $L(\widehat{p}^{(r)} \mid x)$ is monotone increasing in r

Con:

- No guarantee that global (and not local) maximum is obtained
- Convergence can be rather slow

Statistical Inference: Hypothesis testing

1. Definition of hypotheses
2. Choosing the numerical value for the Type I error
3. Selection of a test statistic
4. Determination of the critical value
5. Conduction of the experiment, statistical analysis, decision

Statistical Inference: Definition of hypotheses

Test problem:

$$H_0 : \psi \in \Psi_0 (\subset \Psi) \quad \text{vs.} \quad H_1 : \psi \in \Psi_1 := \Psi \setminus \Psi_0$$

(*null hypothesis* H_0 vs. *alternative hypothesis* H_1)

Examples (Coin tossing):

- $H_0 : p = \frac{1}{2}$ vs. $H_1 : p \neq \frac{1}{2}$ (*two-sided* hypothesis)
- $H_0 : p \leq \frac{1}{2}$ vs. $H_1 : p > \frac{1}{2}$ (*one-sided* hypothesis)

If the hypothesis consists of a single value, it is called a *simple hypothesis*.

Hypotheses consisting of more than a single value are called *composite hypotheses*.

Statistical Inference: Type I and Type II error

Two types of errors:

- Type I error: rejection of H_0 when it is true
- Type II error: acceptance of H_0 when it is false

A procedure frequently adopted is to fix the numerical value α of the Type I error at some low level (e.g. 1% or 5%).

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

Let r denote the number of "heads" in $n = 20$ trials. Possible decision rule:

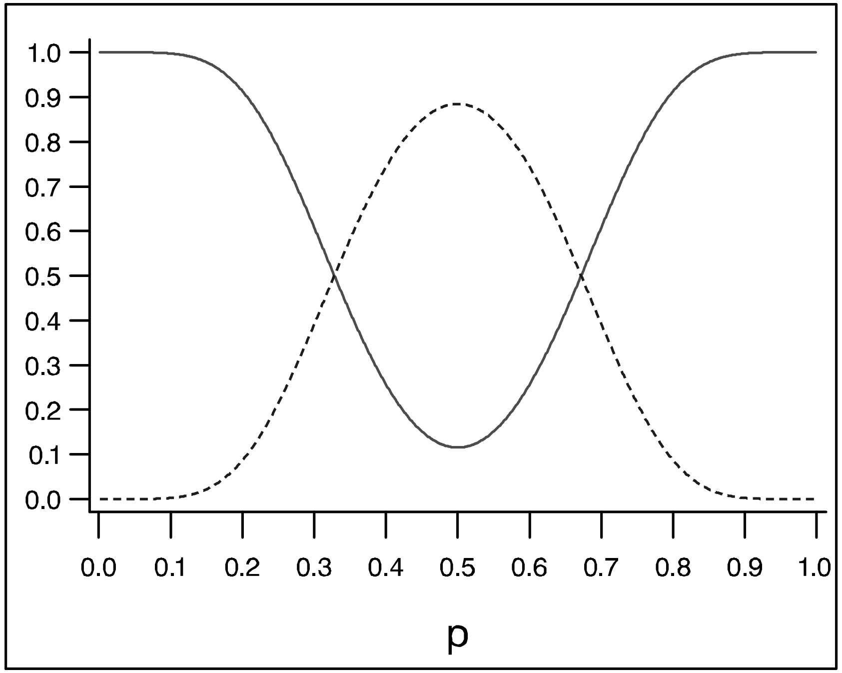
$$r \begin{cases} \leq 6 \quad \text{or} \quad \geq 14 : & \text{reject } H_0 \\ \in [7, 13] : & \text{accept } H_0 \end{cases}$$

Type I error: $\sum_{r=0}^{6} \binom{20}{r} \left(\frac{1}{2}\right)^{20} + \sum_{r=14}^{20} \binom{20}{r} \left(\frac{1}{2}\right)^{20}$

Type II error: $\sum_{r=7}^{13} \binom{20}{r} \cdot p^r \cdot (1-p)^{20-r}$

Statistical Inference: Type I and Type II error

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$



Probability of rejection of H_0 (solid line) and probability of Type II error (dotted line) for the decision rule "Reject H_0 if $r \leq 6$ or $r \geq 14$, otherwise accept H_0".

Statistical Inference: Test statistic

A *test statistic* $T : \mathcal{X} \to \mathbb{R}$ is the quantity calculated from the experimental data whose numerical value leads to acceptance or rejection of the null hypothesis.

Example (Coin tossing):

- $T(x)$: number of heads
- $T(x)$: maximum number of consecutive heads or tails

How to find a “good” test statistic? $\quad\to$ Mathematical statistics

Statistical Inference: Neyman-Pearson Lemma

Consider the test problem of two simple hypotheses, i.e.,

$$H_0 : \psi = \psi_0 \quad \text{vs.} \quad H_1 : \psi = \psi_1.$$

Let $T(x) := L(\psi_1 \mid x)/L(\psi_0 \mid x)$ denote the likelihood ratio. Let $c := \inf\{t : P_{\psi_0}(T > t) \leq \alpha\}$ and let γ satisfy

$$P_{\psi_0}(T > c) + \gamma \cdot P_{\psi_0}(T = c) = \alpha.$$

Now, consider the test which

- rejects H_0 in case of $T(x) > c$,
- accepts H_0 in case of $T(x) < c$,
- rejects H_0 with probability γ in case of $T(x) = c$.

Obviously, the Type I error of this test is equal to α. Further, this test possesses the smallest Type II error probability of all tests of size α.

Example: Neyman-Pearson Lemma

Exercise (Coin tossing):

Suppose a coin is flipped $n = 10$ times. Consider the test problem of two simple hypotheses, i.e.,

$$H_0 : p = p_0 = 0.5 \quad \text{vs.} \quad H_1 : p = p_1 = 0.7.$$

Construct the test of size $\alpha = 0.05$ which possesses the smallest Type II error probability.

(Hint: Show that the likelihood ration $L(p_1 \mid x)/L(p_0 \mid x)$ is increasing in $r = \sum x_i$, i.e., $L(p_1 \mid x)/L(p_0 \mid x) < L(p_1 \mid \tilde{x})/L(p_0 \mid \tilde{x})$ if and only if $\sum x_i < \sum \tilde{x}_i$.)

Statistical Inference: Likelihood ratio test

Consider the general test problem of two hypotheses, i.e.,

$$H_0 : \psi \in \Psi_0 \quad \text{vs.} \quad H_1 : \psi \in \Psi_1 .$$

Let $T(x) := -2 \ln \left(\sup_{\psi \in \Psi_0} L(\psi \mid x) / \sup_{\psi \in \Psi} L(\psi \mid x) \right)$ and consider the test which

- rejects H_0 in case of $T(x) > c$,
- accepts H_0 in case of $T(x) \leq c$.

This test is called the *likelihood ratio test* (LRT).

How to determine c ?

Statistical Inference: Critical value

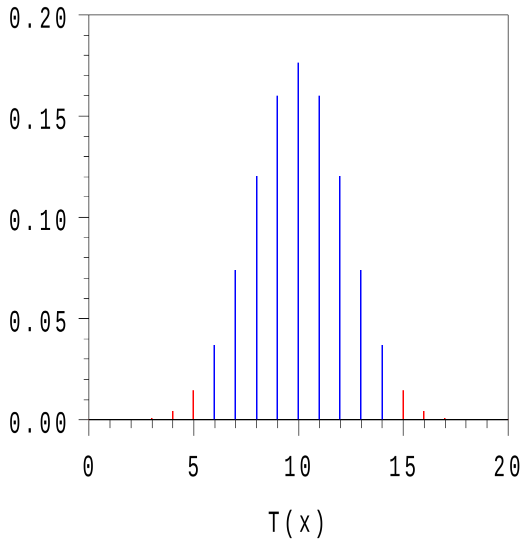
Determination of the critical value requires to calculate the distribution of the test statistic under the null hypothesis. This can be achieved by the following methods:

- Exact calculation
- Approximate calculation
- Simulation (by computer)

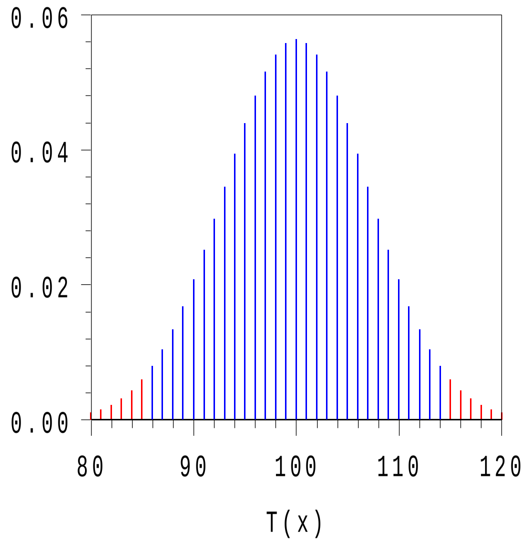
Determination of critical value: Exact calculation

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

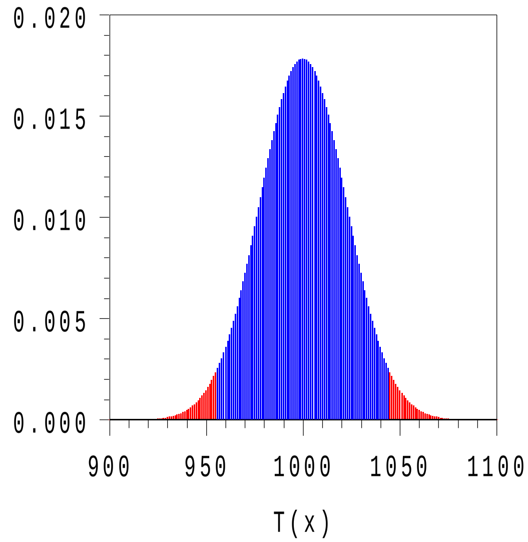
$$T(x) := \sum_{i=1}^{n} x_i$$



$n = 20, \tilde{\alpha} = .0414$



$n = 200, \tilde{\alpha} = .0400$



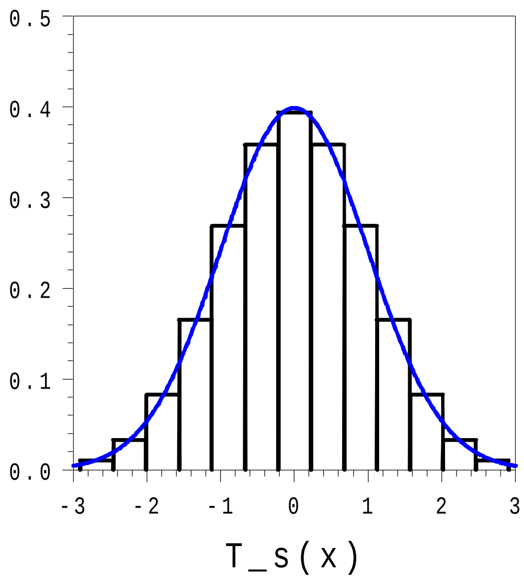
$n = 2000, \tilde{\alpha} = .0466$

Critical values: Approximate calculation

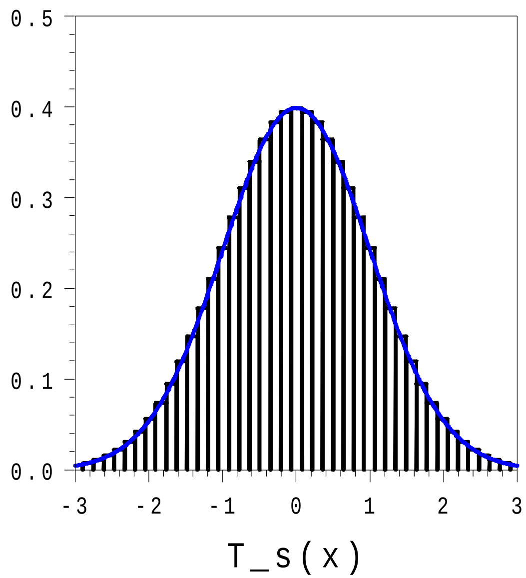
Let $T \sim \mathrm{Bin}(n,p)$ and $T_s := \frac{T - \mathrm{E}T}{\sqrt{\mathrm{Var}T}} = \frac{T - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$.

For sufficiently large n, the distribution of T_s can be approximated by $N(0,1)$.

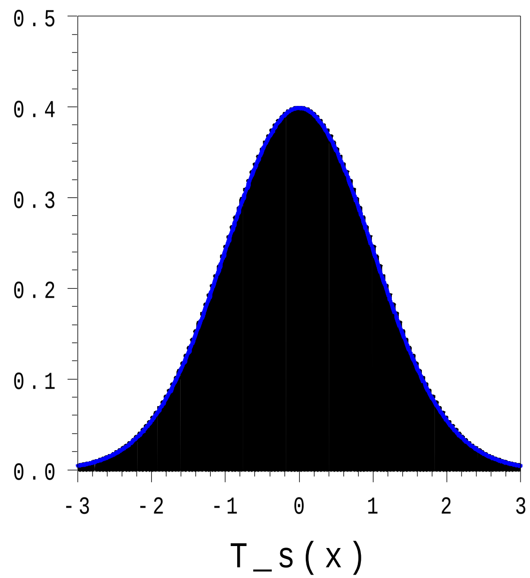
$p = 0.5$:



$n = 20$



$n = 200$



$n = 2000$

Critical values: Approximate calculation

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

Reject H_0 in case that $| T_s(x) | \geq u_{1-\alpha/2}$

$\Leftrightarrow$

Reject H_0 in case that $T(x) \leq n \cdot p - u_{1-\alpha/2} \cdot \sqrt{n \cdot p \cdot (1-p)}$

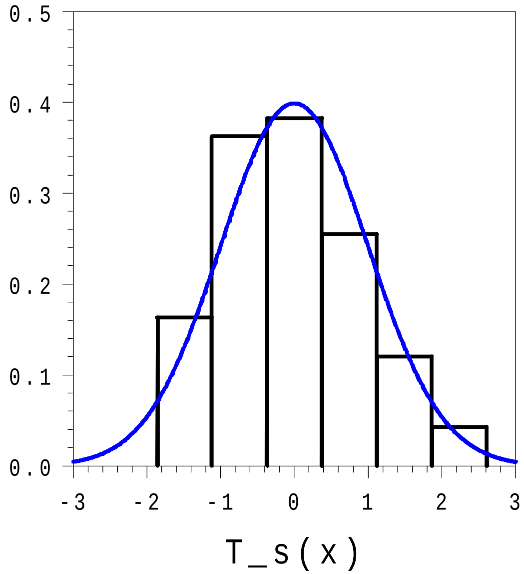
or $T(x) \geq n \cdot p + u_{1-\alpha/2} \cdot \sqrt{n \cdot p \cdot (1-p)}$.

$\alpha = 0.05$:

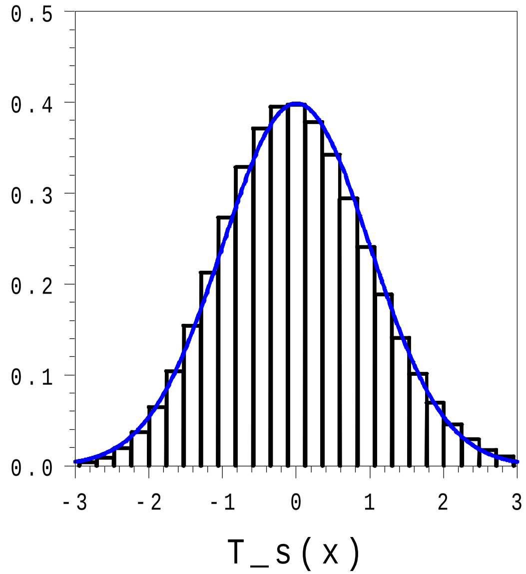
n	true Type I error rate
20	.0414
200	.0560
2000	.0517

Critical values: Approximate calculation

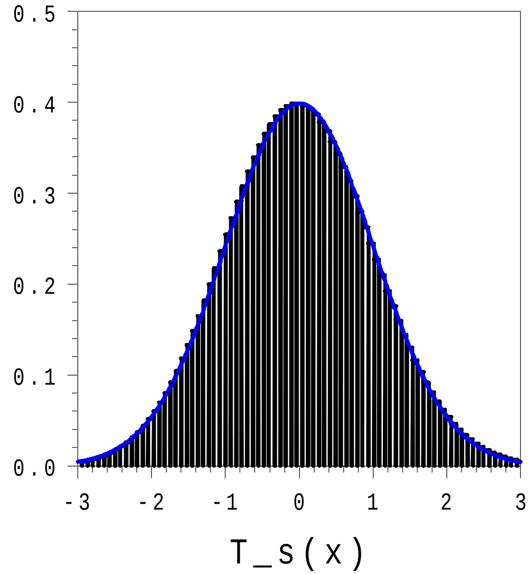
$p = 0.1$:



$n = 20$



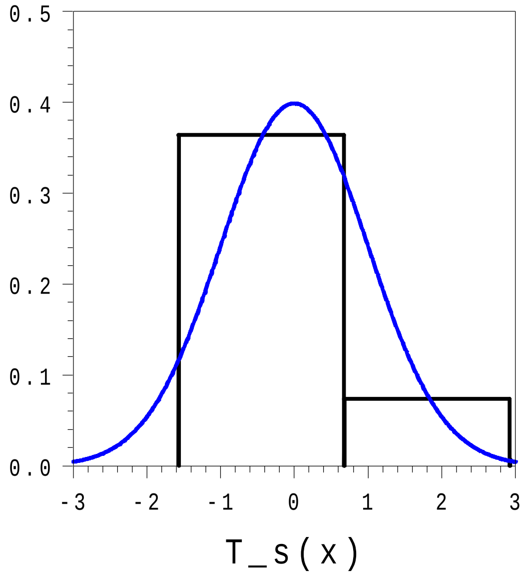
$n = 200$



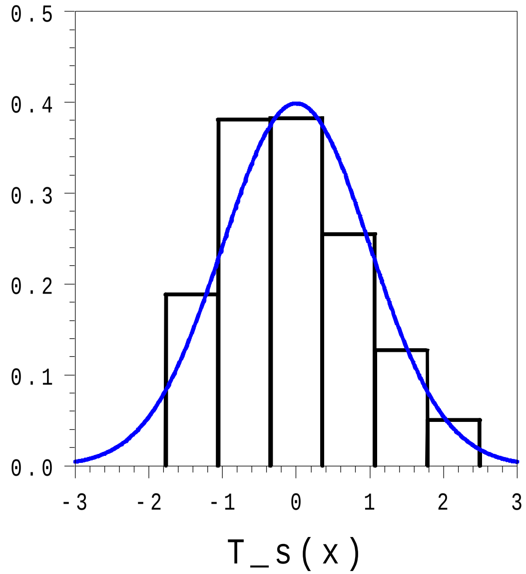
$n = 2000$

Critical values: Approximate calculation

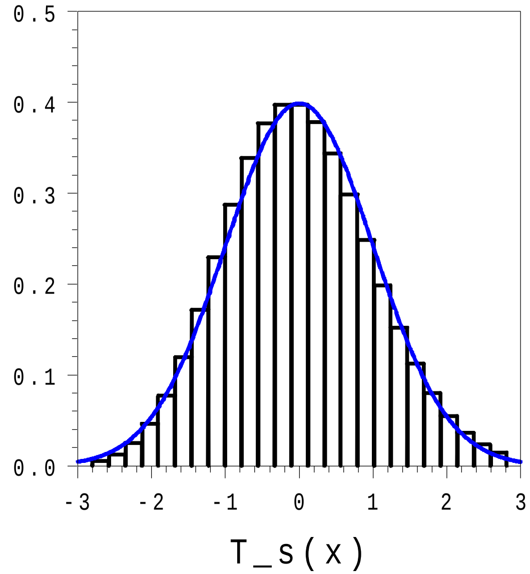
$p = 0.01$:



$n = 20$



$n = 200$



$n = 2000$

Determination of critical value: Simulation

1. Draw x (under the null hypothesis).
2. Calculate $T(x)$.
3. Perform m replicates of steps 1 and 2, and calculate the empirical null distribution of $T(x)$.
4. Use this empirical distribution to obtain critical values.

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

$n = 20$, $m = 100,000$:

r	$P(T(x) = r)$	$\widehat{P}(T(x) = r)$		r	$P(T(x) = r)$	$\widehat{P}(T(x) = r)$
0	.00000	.00000		16	.00462	.00456
1	.00002	.00000		17	.00109	.00135
2	.00018	.00016		18	.00018	.00012
3	.00109	.00080		19	.00002	.00002
4	.00462	.00458		20	.00000	.00000

P-value

Instead of calculation of critical values:

The probability (under the null hypothesis) of obtaining the observed value of the test statistic or a more extreme value is called P-*value*. If the P-value is $\leq$ than the chosen type I error rate α, the null hypothesis is rejected.

Example (Coin tossing): $H_0 : p = 1/2$ vs. $H_1 : p \neq 1/2$

Assume that the observed number of "heads" out of $n = 20$ trials is $r = 18$.

Then, the null probability

$$P(T \geq 18) = \sum_{r=18}^{20} \binom{20}{r} \cdot (1/2)^{20} = .0002.$$

Since a two-sided alternative H_1 is considered, $P(T \leq 2)$ has to be added.

$\Rightarrow$ P-value corresponding to $r = 18$ is $P(T \geq 18) + P(T \leq 2) = .0004$.