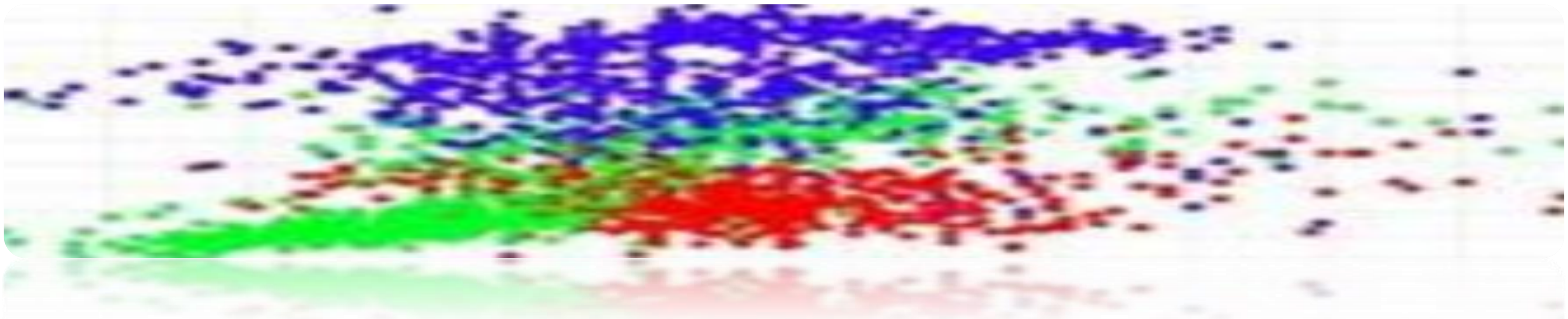


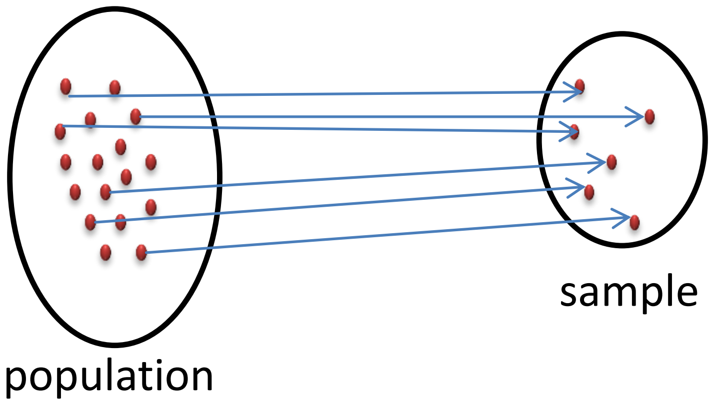
BASICS OF STATISTICS

Outline

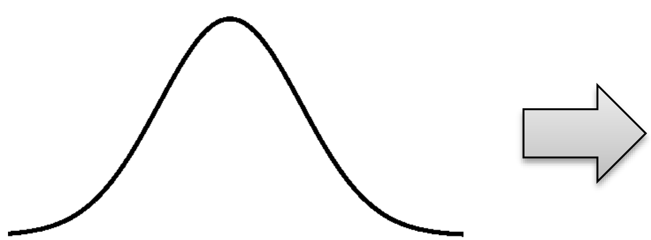
- Sampling
- Estimators, bias, consistency, and mean squared error
- Law of Large Numbers
- Central Limit Theorem
- Hypothesis testing

What is statistics about?

- Statistics is concerned with data that are subject to random variations
  - Collecting data through sampling,



  - Summarizing and analyzing data by estimating the parameters of the underlying distribution(s)



X: height of a person

1.75 1.92 1.69 1.80 ....

Estimate mean: $\hat{\mu} \approx E(X)$
Estimate variance: $\hat{\sigma}^2 \approx Var(X)$

Sampling

Population

Convenience sample (biased)

Only data that is easy to accessible is sampled

Systematic sample (biased)

Ordered sampling: randomly select first element and then select every k'th element

…

Stratified sample (unbiased)

When there is variance in subpopulations, it is better to sample each population Independently (sample size should reflect the proportion of the subpopulation)

Random sample (unbiased)

Every element has equal probability of being selected

Estimators

- Definition: An **estimator** is a function that uses input from the sample space to estimate a parameter of the underlying data distribution

- Examples: Let $x_1, \dots, x_n$ be the values of i.i.d. random variables X_i

  - **Empirical mean** and the **sample mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  - **Empirical variance:** $S_{em}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$

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

Covariance estimators and correlation

- Let $(x_1, y_1), \ldots, (x_n, y_n)$ be samples of i.i.d. random variables X_i, Y_i
  - Empirical covariance: $\hat{C}_{em} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$
  - Sample covariance: $\hat{C} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$
  - Correlation: $r = \frac{\hat{C}}{s_X s_Y}$

For linear dependency between two variables, e.g., $Y = aX + b$:

$$r = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}$$

Empirical distribution function and empirical median

- Let $x_1, \dots, x_n$ be the values of i.i.d. random variables X_i

  - Empirical distribution function: $\hat{F}_{X_{i:n}}(x) = \frac{1}{n}\sum_{i=1}^{n}[\![x_i \leq x]\!]$

    where $[\![x_i \leq x]\!] := \begin{cases} 1, x_i \leq x \\ 0, x_i > x \end{cases}$ is called the indicator function

  - Empirical median $\hat{x}_{med}$ is defined as $\hat{F}_{X_{i:n}}(\hat{x}_{med}) = \frac{1}{2}$, that is, for ordered $x_{i_1} \leq \cdots \leq x_{i_n}$:

$$\hat{x}_{med} = \begin{cases} x_{\left(i_{(n+1)/2}\right)} & for\ odd\ n \\ \frac{\left(x_{i_{n/2}} + x_{i_{(n+2)/2}}\right)}{2} & for\ even\ n \end{cases}$$

Example

- What is the expected life time of a specific electronic device (in months)?
- Random variable X:= life time in # months
- Random sample:

  $x_1 = 38,\ x_2 = 33,\ x_3 = 35,\ x_4 = 32,\ x_5 = 9,\ x_6 = 36,\ x_7 = 31,$
  $x_8 = 37,\ x_9 = 22,\ x_{10} = 40,\ x_{11} = 30$

  - Empirical mean: $\bar{X} = \frac{1}{11}\sum_{i=1}^{11} x_i \approx 31.2$
  - Empirical median: 33
  - Empirical variance: $S_{em}^2 = \frac{1}{11}\sum_{i=1}^{11}(x_i - \bar{X})^2 \approx 70.69$
  - Sample variance: $S^2 = \frac{1}{10}\sum_{i=1}^{11}(x_i - \bar{X})^2 \approx 77.76$

Estimators

- How “good” is an estimator?
  - How well does it approximate the true parameter on average?
  - Can it yield the true parameter with more and more data?
  - What is the variance of the estimator?

- Definition: An estimator $\hat{\gamma}$ is unbiased if its expected value $E(\hat{\gamma})$ is equal to the true value of the parameter γ it estimates, i.e., $E(\hat{\gamma}) = \gamma$, otherwise $\hat{\gamma}$ is biased with squared bias $(E(\hat{\gamma}) - \gamma)^2$

- Definition: An estimator $\hat{\gamma}$ derived from n values of i.i.d. random variables X_i is consistent if $\lim_{n\to\infty} P(|\hat{\gamma} - \gamma| > \varepsilon) = 0$ for all ε>0

Mean and variance estimators

- Let $x_1, \dots, x_n$ be the values of i.i.d. random variables X_i
- Theorem
  - The empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an **unbiased consistent** estimator of the true mean $E(X)$
  - The empirical variance $S_{em}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$ is a **biased consistent** estimator of the true variance $Var(X)$, it can be shown that $E(S_{em}^2) = \frac{n-1}{n}Var(X)$
  - The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$ is an **unbiased consistent** estimator of the true variance $Var(X)$
  - The sample covariance $\hat{C} = \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{X})(y_i - \bar{Y})$ is an **unbiased consistent** estimator of $Cov(X)$
  - The empirical distribution function $\hat{F}_{X_{i:n}}(x) = \frac{1}{n}\sum_{i=1}^{n}[\![x_i \le x]\!]$ is an **unbiased consistent** estimator of the true cumulative distribution F_X

Law of Large Numbers

- Let $x_1, x_2, \dots, x_n$ be a random sample from a distr. $f_X(x)$ and $\bar{X} = \sum_i \frac{X_i}{n}$

- Weak law of large numbers (weak consistency of the empirical mean)
  - $\lim_{n \to \infty} P(|\bar{X} - E(X)| > \varepsilon) = 0$ for all ε>0
  - Sample average converges in probability towards the mean of the distr. of X

- Strong law of large numbers (strong consistency of the empirical mean)
  - $P(\lim_{n \to \infty} |\bar{X} - E(X)| > \varepsilon) = 0$ for all ε>0
  - Sample average converges almost surely towards the mean of the distr. of X

Best estimators

- Definition: An unbiased estimator $\hat{\gamma}$ is the **best estimator** of the true parameter γ if it has lowest variance among all other unbiased estimators, i.e., for all unbiased estimators $\hat{\gamma}'$ of γ: $Var(\hat{\gamma}) \leq Var(\hat{\gamma}')$

- The **mean squared error** between an estimator $\hat{\gamma}$ and γ is:

$$mse(\hat{\gamma} - \gamma) = E\left((\hat{\gamma} - \gamma)^2\right) = Var(\hat{\gamma}) + Bias(\hat{\gamma})^2$$

Because:

$$Var(\hat{\gamma}) = Var(\hat{\gamma} - \gamma) = E\left((\hat{\gamma} - \gamma)^2\right) - E^2(\hat{\gamma} - \gamma)$$

$$Bias(\hat{\gamma}) = E(\hat{\gamma}) - \gamma = E(\hat{\gamma}) - E(\gamma) = E(\hat{\gamma} - \gamma)$$

- Notes
  - The sample mean is the best estimator of the true mean for many useful distributions
  - The sample variance is the best estimator of the true variance for normally distributed data

Usefulness of estimators

- How useful is an estimator for the understanding of the underlying distribution?

- It depends on the distribution!

- Example
  - Random variable X:= yearly income in \$1000
  - Random sample:

$$x_1 = 58;\ x_2 = 74;\ x_3 = 69;\ x_4 = 81;\ x_5 = 64;\ x_6 = 120;\ x_7 = 55;$$
$$x_8 = 71;\ x_9 = 77;\ x_{10} = 65;\ x_{11} = 23{,}000 \quad \Rightarrow \quad \bar{X} \approx 2{,}158$$

  - Empirical median is more insightful in this case

Other useful estimators

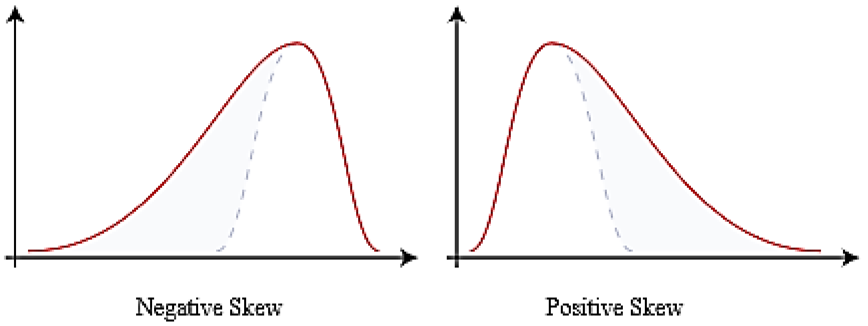
- Maximum Likelihood Estimator, i.e., $argmax_\theta P(x_1, \dots, x_n | \theta)$
  - Consistent
  - Asymptotically normal
  - Asymptotically optimal, i.e., with smallest variance

- $Min(x_1, \dots, x_n)$

- $Max(x_1, \dots, x_n)$

- Empirical skewness

$$Sk = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2\right)^{3/2}}$$



Source: Wikipedia

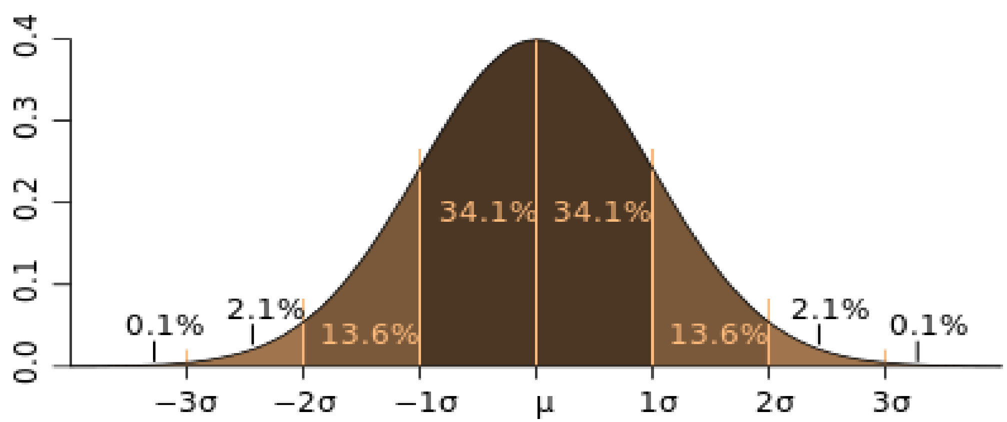
Recap of the normal distribution

- X is normally distributed $\Leftrightarrow$ $X \sim N(\mu, \sigma^2) \Leftrightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  μ: *mean*, σ: *standard deviation*

- Standard normal distribution: $N(0,1)$

- Cumulative distribution of $N(0,1)$: $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}} \, dx$



- Theorem: If $X \sim N(\mu, \sigma^2)$ then $Y := \frac{(X-\mu)}{\sigma} \sim N(0,1)$

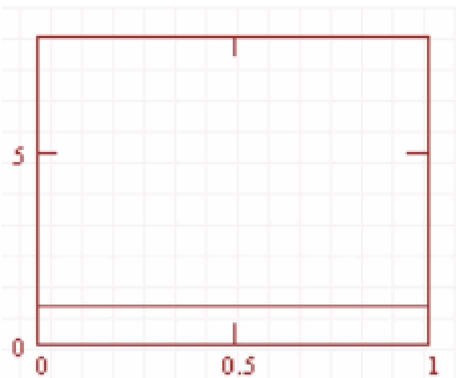
Central Limit Theorem

- **Central Limit Theorem**: Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a distr. with mean μ and finite non-zero variance σ^2. The cdf of the random variable $Z := \sum_i X_i$ converges to the cdf of the normal distribution $N(n\mu, n\sigma^2)$. That is:

$$\lim_{n\to\infty} P\left(a \le \frac{Z - n\mu}{\sqrt{n}\sigma} \le b\right) = \Phi(b) - \Phi(a)$$

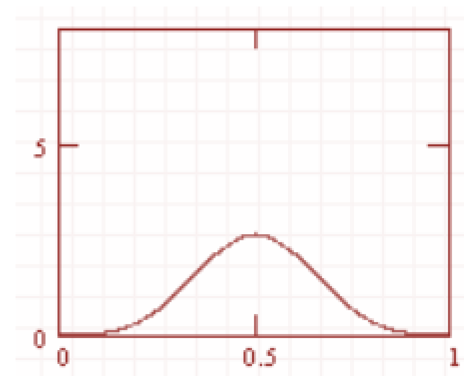
- **Corollary:** The cdf of $Z := \frac{1}{n}\sum_i X_i$ converges to the cdf of $N\left(\mu, \frac{\sigma^2}{n}\right)$

Central Limit Theorem: Example



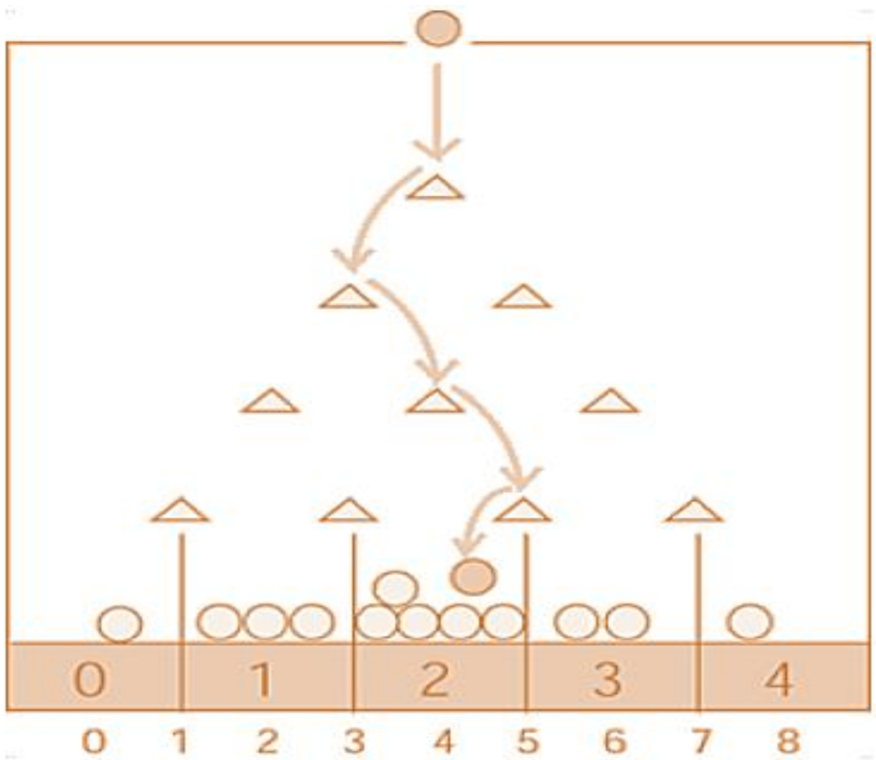
$f_X(x)$ is uniform

Avg. of X_1, X_2 sampled repeatedly from $f_X(x)$



Avg. of X_1, X_2, X_3, X_4 sampled repeatedly from $f_X(x)$

Galton Machine



Empirical evidence for the Central Limit Theorem (by considering sequences of i.i.d. Bernoulli variables) and for the Law of Large Numbers (by considering random samples from a Binomial distribution)

Hypothesis testing

- Example hypotheses:
  - Sample originates from normal distribution
  - Two random variables are independent
  - Sample is Bernoulli distributed with p=0.5

- Goal: Falsification of hypothesis by lack of statistical evidence

  - Hypothesis to be falsified: H_0 (null hypothesis)
  - Counter hypothesis: H_1
  - Test region R from cdf of test variable X
    - $X \in R \Rightarrow$ reject H_0
    - $X \notin R \Rightarrow$ retain H_0

	Retain H_0	Reject H_0
H_0 true	ok	Type I error
H_1 true	Type II error	ok

Hypothesis testing: Example

Assume average IQ
of students is 100
$H_0: \mu = 100$

Run IQ test
on sample

IQ_1 IQ_2 IQ_3 IQ_4 IQ_5 IQ_6 IQ_7 IQ_8

$\overline{IQ} = 115$

Is this likely given $\mu = 100$?

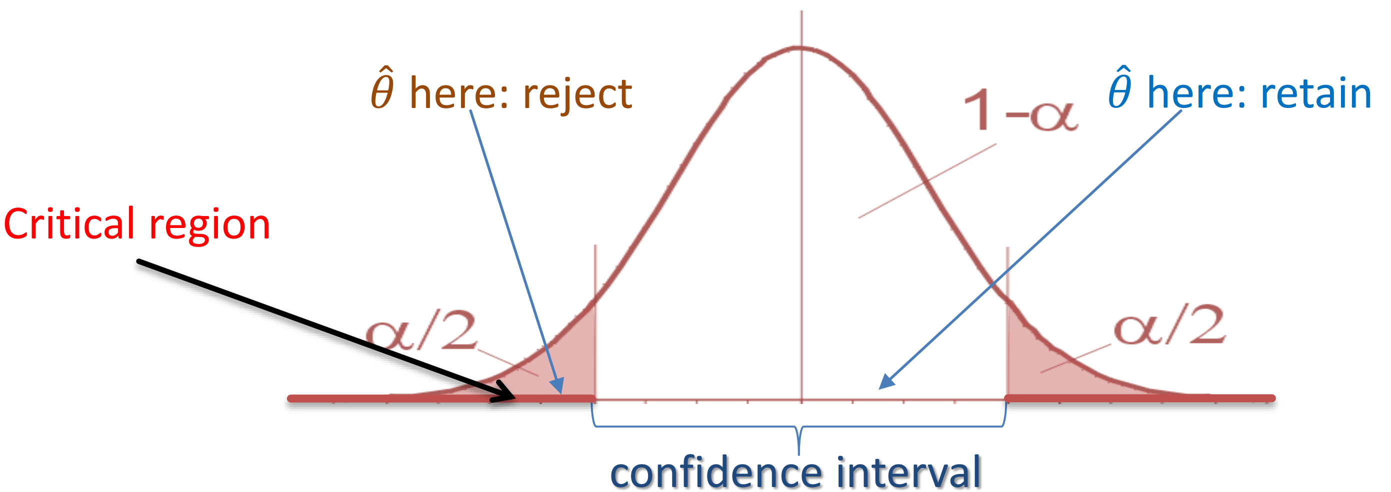
If yes retain H_0 else reject

Confidence interval and confidence level

- How well is a parameter estimated
  - Consider estimator $\hat{\theta}$ for parameter θ
  - How well does $\hat{\theta}$ represent θ?

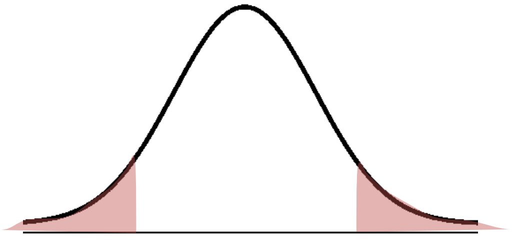
$$P\left(\hat{\theta} - c \leq \theta \leq \hat{\theta} + c\right) = 1 - \alpha$$

- Definitions
  - The interval $[\hat{\theta} - c, \hat{\theta} + c]$ is the **confidence interval**
  - The value $1 - \alpha$ is the **confidence level**
  - α is the **significance level** (typically: 0.01, 0.05, 0.1)

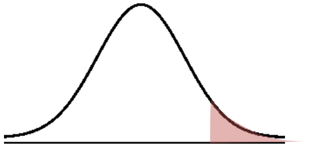


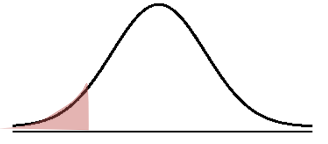
One sided and two-sided tests

- 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 either of these forms
  - $H_0: \theta \leq \theta_0 \; vs. H_1: \theta > \theta_0$
  - $H_0: \theta \geq \theta_0 \; vs. H_1: \theta < \theta_0$





is called a one-sided test

Unknown mean and known variance

- Consider i.i.d. random variables $X_1, \dots, X_n,\ n \gg 1$, from a distribution with unknown, non-zero mean μ and known finite variance σ^2.
  - We know $\bar{X} = \frac{1}{n}\sum_i X_i$ is approximately normally distributed with $N(\mu, \frac{\sigma^2}{n})$
  - We also know that $Y = \frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \sim N(0,1)$

$$P\left(-z \leq \frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \leq z\right) = \Phi(z) - \Phi(-z) = P\left(\bar{X} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z\sigma}{\sqrt{n}}\right)$$

$\Rightarrow$ For confidence interval $[\bar{X} - c, \bar{X} + c]$ set $z \coloneqq \frac{c\sqrt{n}}{\sigma}$ and look up $\Phi(z)$

$\Rightarrow$ For confidence level $1 - \alpha$, and a proposed value for μ, reject null hypothesis if $|Y| > \Phi^{-1}(1 - \alpha/2)$

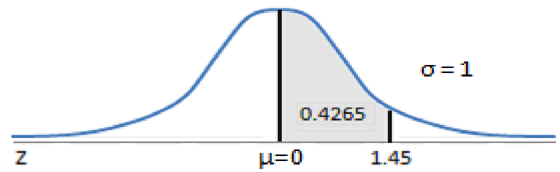
- Definition: The p-value is minimal significance level at which H_0 can be rejected

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Z-score table

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score.
For example, when Z score = 1.45 the area = 0.4265.



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995

Wald Test

- For a parameter $\hat{\theta}$ derived from a sample and a proposed parameter θ, we can test

$$H_0: \hat{\theta} = \theta \text{ vs. } H_1: \hat{\theta} \neq \theta$$

  - $s = \sqrt{Var(\hat{\theta})}$ is called the standard error and $Var(\hat{\theta})$ is the sample variance
  - Test variable $W := \frac{\hat{\theta} - \theta}{s}$ is approximately $N(0,1)$-distributed (i.e., distribution of W converges to $N(0,1)$ for growing sample size)

→ Reject H_0 at level α when $|W| > \Phi^{-1}(1 - \alpha/2)$

- Example
  - $\hat{\theta}$: Average increase of height of men compared to height of women
  - Proposed parameter $\theta = 0$

Wald test: Example

- What is the expected life time of a specific electronic device (in months)?
- Random variable X:= life time in # months
- Random sample:

$x_1 = 38,\ x_2 = 33,\ x_3 = 35,\ x_4 = 32,\ x_5 = 9,\ x_6 = 36,\ x_7 = 31,$
$x_8 = 37,\ x_9 = 22,\ x_{10} = 40,\ x_{11} = 30$

  - Empirical mean: $\bar{X} = \frac{1}{11}\sum_{i=1}^{11} x_i \approx 31.2$
  - Sample variance: $S^2 = \frac{1}{10}\sum_{i=1}^{11}(x_i - \bar{X})^2 \approx 77.76$

- Hypothesis I: Devices have a life time of around 2 years

$$W := \frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \approx 0{,}82 < 1.96 \text{ (for significance level 0.05)}$$

- Hypothesis II: Devices have a life time of around 1 year

$$W := \frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \approx 2{,}177 > 1.96 \text{ (for significance level 0.05)}$$

Example: Probability of heads

- H_0: coin has head probability $p = p_0$
- X: test variable representing #heads in n tosses
- We know that approximately $X \sim N(pn, p(1-p)n)$

- $Y := \frac{(X-pn)}{\sqrt{p(1-p)n}} \sim N(0,1) \Rightarrow$ reject H_0 at level α $(= 0.05)$ if

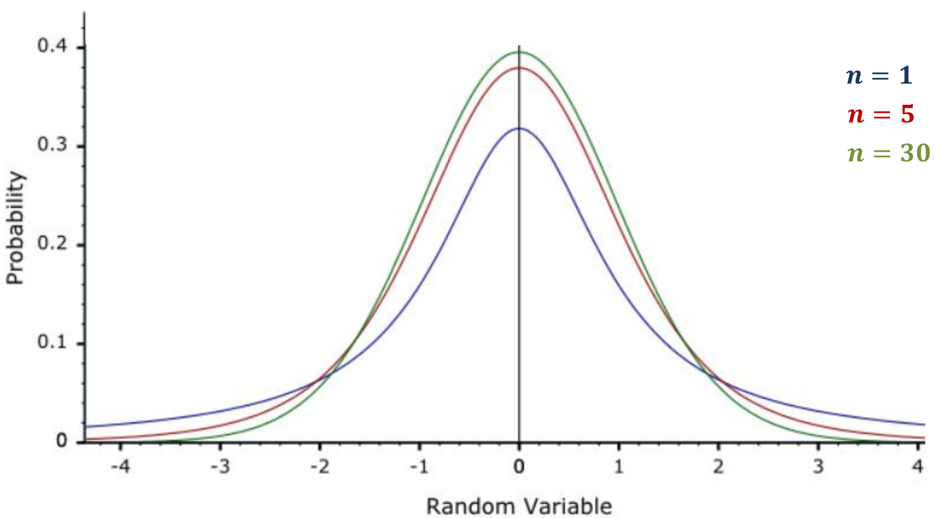
$$Y > \Phi^{-1}(1-\alpha/2) \text{ or } Y < \Phi^{-1}(\alpha/2) \Leftrightarrow |Y| > \Phi^{-1}(1-\alpha/2)$$

t-Test for unknown mean and unknown variance

- Consider i.i.d. random variables $X_1, \ldots, X_n,\ n \gg 1$, from a distribution with **unknown**, non-zero mean μ and **unknown** variance
- Let s^2 be the sample variance. $Y := \frac{(\bar{X}-\mu)\sqrt{n}}{s}$ has a **Student's *t* distribution** with $n-1$ degrees of freedom
  - With analogous derivation as before:

$$P\left(\bar{X} - \frac{t_{n-1,1-\alpha/2}s}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{t_{n-1,1-\alpha/2}s}{\sqrt{n}}\right) = 1-\alpha$$

$\Rightarrow$ For proposed μ and significance level α, reject null hypothesis if $|Y| > t_{n-1,1-\alpha/2}$



t-Test in practice

- Compare two prediction algorithms A and A' based on performance on k labeled datasets
- Let $e_1, \ldots, e_k$ and $e_1', \ldots, e_k'$ be the error values (or any performance values), respectively
- Are the error means any different?

- Fact: $\bar{e}$ and $\overline{e'}$ are approximately normally distributed, but we neither know the means nor the variances

- Since σ_e and σ_e' are unknown, we need to use ***t*-distribution** with k-1 degrees of freedom to estimate how close μ_e and μ_e' are (H_0: $\mu_e = \mu_e'$)
- $\bar{d} = \bar{e} - \overline{e'}$ is also t-distributed, with k-1 degrees of freedom

$$\Rightarrow H_0: \bar{d} = 0 \text{ and } Y := \frac{(\bar{d}-0)\sqrt{k}}{s_d} \text{ is the t-statistics}$$

- Use t-distribution table to determine the $t_{k-1,1-\alpha/2}$ score
- If $t_{k-1,1-\alpha/2} < |Y|$ reject H_0 otherwise retain it

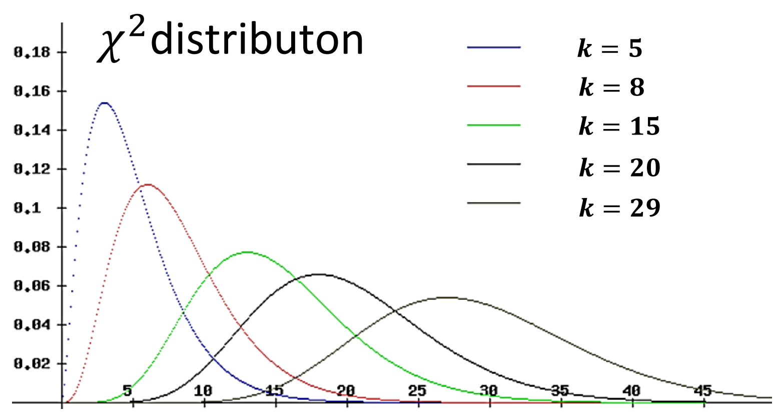
Chi-Square Goodness-of-Fit-Test

- Given sample $x_1, \dots, x_n$ of i.i.d. random variables X_i and absolute frequencies $h_1, \dots, h_k$ of class $c_j, 1 \leq j \leq k$, we can test

  - H_0: X_i follow a proposed discrete distribution

  - $Z_k := \frac{\sum_{j=1}^{k} \left(h_j - E(h_j)\right)^2}{E(h_j)}$, with $E(h_j)$ being the expected frequency of class c_j according to the proposed distribution, is χ^2-distributed with k-1 degrees of freedom

    → Reject H_0 at test level α (e.g. 0.05) if $Z_k > \chi^2_{k-1,1-\alpha}$



IT Systems Engineering | Universität Potsdam

Chi square distribution table

d.f.	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	32.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	32.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3

Chi-Square independence test

- r = number of columns
- m = number of rows
- n_{ij} = Actual number in cell$_{ij}$
- n^*_{ij}= Expected number in cell$_{ij}$
- $(r-1)(m-1)$ = degrees of freedom

	Feature Y						Sum Σ
Feature X	1	2	…	k	…	r	$n_{j.}$
1	n_{11}	n_{12}	…	n_{1k}	…	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	…	n_{2k}	…	n_{2r}	$n_{2.}$
…	…	…	…	…	…	…	…
j	…	…	…	n_{jk}	…	…	$n_{j.}$
…	…	…	…	…	…	…	…
m	n_{m1}	n_{m2}	…	n_{mk}	…	n_{mr}	$n_{m.}$
Sum Σ	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	…	$n_{.r}$	n

$$n^*_{jk} = \frac{n_{j.} \cdot n_{.k}}{n}$$

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}}$$

→ Reject H_0 at test level α (e.g. 0.05) if $\chi^2 > \chi^2_{(r-1)(m-1),1-\alpha}$

General recipe for hypothesis testing

- Formulate null hypothesis
- Define corresponding random variable for the test
- Turn the variable into a $N(0,1)$-distributed variable, or a t-statistics, or a χ^2-statistics, …
- Test whether the new statistics lies in the critical region of the underlying distribution