LStat
Leuven Statistics Research Centre

Fundamental Concepts of Statistics

Chapter 5: Sample statistics & Limit Theorems

Martial Luyts & Clement Cerovecki

Catholic University of Leuven, Belgium

martial.luyts@kuleuven.be

KU LEUVEN

LEUVEN STATISTICS
RESEARCH CENTRE

Contents

Part 1:

Sample Statistics

1. Introductory material

- **Fact:** The distribution of a random variable (rv) X is unknown.
- Given the context of the underlying random phenomenon, we can say that F_X belongs to some specified family $\mathcal{F}$ of distribution functions.
- **Question:** Which member of $\mathcal{F}$ is the true distribution function of X?
- To answer this question, one distinguishes between (2) parametric & (3) non-parametric (1) statistical models.
- In what follows, we first explain these concepts!

TERMINOLOGY:

1. **Statistical model**

   $=$ a class of distributions used to describe a rv X.

   Examples:

   - $M_1 = \{N(\mu, \sigma^2) | \mu \in \boldsymbol{R}, \sigma > 0\}$, two-parameter normal family
   - $M_2 = \{N(\mu_0, \sigma^2) | \sigma > 0\}$, one-parameter normal family with given mean
   - $M_3 = \{$the set of symmetric bell-shaped distributions$\}$
   - $M_4 = \{F : \; F$ is a continuous cdf$\}$

2. Parametric model

$=$ a statistical model of probability density functions (pdf's) which depend only on a finite-dimensional parameter:

$$M = \{f(x,\theta) | \theta = (\theta_1, \ldots, \theta_k) \in \Theta \subset \boldsymbol{R}^k\}$$

where f is a pdf depending on a $k-$dimensional parameter θ. The **parameter space** Θ is the set of possible parameter values in the model.

Our examples: M_1 & M_2 are parametric models:

- $M_1: \;\theta = (\mu, \sigma),\; \Theta = \{(\mu, \sigma) : \mu \in \boldsymbol{R}, \sigma > 0\} \subset \boldsymbol{R}^2$
- $M_2: \;\theta = \sigma,\; \Theta = \{\sigma \in \boldsymbol{R} : \sigma > 0\} \subset \boldsymbol{R}$

3. **Non-parametric model**

$=$ a model which cannot be described by a finite-dimensional parameter.

Our examples: M_3 & M_4 are non-parametric models

- In this course, we will focus on **parametric statistical models**, hence each member of the family $\mathcal{F}$ can be specified by parameter θ of a finite dimension k.
- **Problem:** Population parameter θ is unknown.
- **Question:** Is there a way to derive information from it?

- **Answer:** Based on a given sample $X_1, \ldots, X_n$ from X, estimators for θ can be derived. Sample surveys are used to obtain information about a large population by examining only a small fraction of that population.

- One can have several estimators for the same parameter. This raises the question whether one of the estimators should be preferred, and why.

- The theory of (sample) statistics based on parametric models is called **parametric statistics**. **Non-parametric statistics (distribution-free statistics)** is the theory of (sample) statistics which does not assume a parametric model.

- In view of a sound statistical theory we make a distinction between data sample and mathematical sample, and introduce the following concepts:

Population - data sample - mathematical sample

Parameter - data statistic - sample statistic

- **Example:** Distribution of the length in a population

  Purpose: Obtain relevant and reliable information on the distribution of length of individuals in a population.

  - **Random variable**:
    X = Length of a randomly chosen individual in the population (cm)
  - **Statistical model**
    Model distribution used for X: $X \sim N(\mu, \sigma^2)$
  - **Population parameters**:

    $\mu = E(X)$ population mean (mean length)

    $\sigma^2 = Var(X)$ population variance

  Commonly we have to our disposal a set of n observations $x_1, \ldots, x_n$ that are n realizations of X. These observations are often the results obtained from n independent repetitions of same random experiment, where each repetition of experiment was carried out under same circumstances.

An observed sample of size n: data (observations) $x_1, \ldots, x_n$

- **Observed sample statistics**

  $\bar{x} = \frac{\sum x_i}{n}$ observed sample mean

  $\tilde{s}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ observed sample variance

- New sample provides new data and new data statistics.

  $\Rightarrow$ values $x_1, \ldots, x_n$ can be considered as realizations of n independent **random variables** $X_1, \ldots, X_n$ following the same statistical model (i.e. having same distribution as population rv X).
  Random Sample of size n from X is n-tuple $(X_1, \ldots, X_n)$ of independent rv's all having same distribution as X. $X_1, \ldots, X_n$ are iid (same distribution as X).

- Also the values $\bar{x}$ and $\tilde{s}^2$ are realisations of $\bar{X}$ and $\tilde{S}^2$.

  $\overline{X} = \frac{\sum X_i}{n}$ sample mean

  $\tilde{S}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$ sample variance

2. Strategy for the study of the sample statistic distribution

- Suppose we have n observations $x_1, \ldots, x_n$.

- Let T be given function (not depending on any unknown parameter) defined on $\boldsymbol{R}^n$ and consider $T(x_1, \ldots, x_n)$.

- Definition:

  If the function T is such that $T(X_1, \ldots, X_n)$ is a rv, then $T = T(X_1, \ldots, X_n)$ is a **statistic**.

- Often, a statistical problem reduces to finding the distribution of this statistic.

- Under a general population distribution one usually tries to obtain:

1. the values of $E(T)$ and $Var(T)$ without assumptions on the population distribution, except mild conditions such as the existence of the mean $\mu = E(X)$ and/or second moment $E(X^2)$

2. an approximate distribution
   - asymptotics (asymptotic value and distribution) for large sample size (limit theorems for $n \to \infty$):
   - the distribution of a slightly modified statistic

3. improving approximate distributions

4. the exact distribution of T, under a specific family of population distributions, e.g. under the normal model $X \sim N(\mu, \sigma^2)$

Population	**Data sample (observed)**	**(Mathematical) sample**
Population rv X: rv under investigation. **Population distribution**: distribution of X (i.e. F_X, f_X or M_X).	**Data-sample**: sequence of random observations $x_1, \ldots, x_n$ for X, thus numbers.	**Mathematical sample** of size n from X: sequence of rv's $X_1, \ldots, X_n$ iid as X.
Population parameter: number $\theta = g(F_X)$ function of distribution of X, usually unknown.	**Data statistic**: number $t = T(x_1, \ldots, x_n)$ function of the data ⇒ estimate for θ.	**Sample statistic:** rv $T = T(X_1, \ldots, X_n)$ function of the sample ⇒ estimator for θ.
Problem: obtain information on distribution of X (that is on F_X or f_X, e.g. on the value of a population parameter θ).	Data provide information. Data statistics are used to estimate population parameters.	accuracy/efficiency of an estimate is described by the distribution of the corresponding sample statistic (sampling distribution).

3. Sample statistics

Given iid sample $X_1, \ldots, X_n$ from population X.

Parameter	Corresponding sample statistic
$\mu = E(X)$	$\overline{X} = \frac{\sum X_i}{n}$
$\sigma^2 = E[(X-\mu)^2]$	$V^2 = \frac{\sum (X_i - \mu)^2}{n}$
	$\tilde{S}^2 = \frac{\sum (X_i - \overline{X})^2}{n}, \; S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{n}{n-1}\tilde{S}^2$
$\alpha_k = E(X^k)$	$a_k = \frac{\sum X_i^k}{n}$
$\mu_k = E[(X-\mu)^k]$	$b_k = \frac{\sum (X_i - \mu)^k}{n}, \quad m_k = \frac{\sum (X_i - \overline{X})^k}{n}$

4. Moments of sample statistics

- The sample moments are estimators of the corresponding population moments
- **Question:** But what can we say about their qualities (**properties**)?
- Of particular interest are the sample mean and the sample variance!
- **Reminder:** Consider an iid sample $X_1, \ldots, X_n$ from a population X with mean $E(X) = \mu$ and variance $Var(X) = \sigma^2$, and let

$$\overline{X} = \frac{\sum X_i}{n}, \; V^2 = \frac{\sum (X_i - \mu)^2}{n}, \; \tilde{S}^2 = \frac{\sum (X_i - \bar{X})^2}{n}, \; S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1},$$

respectively called the sample mean, the sample variance with known mean, the sample variance with weight n, the sample variance with weight $n-1$ (or bias corrected sample variance).

SOME PROPERTIES:

- **Linkage:** Variation decomposition theorem: SS lemma

$$\sum_i (X_i - \mu)^2 = \sum_i \left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2 .$$

or, equivalently,

$$V^2 = \tilde{S}^2 + \left(\overline{X} - \mu\right)^2 .$$

- **Moments:**

1. For the sample mean $\overline{X}$:

$$E(\overline{X}) = \mu, \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

2. For the sample variance with known mean:

$$E(V^2) = \sigma^2, \qquad Var(V^2) = \frac{\mu_4 - \sigma^4}{n}$$

   where $\mu_k = E(X - \mu)^k$.

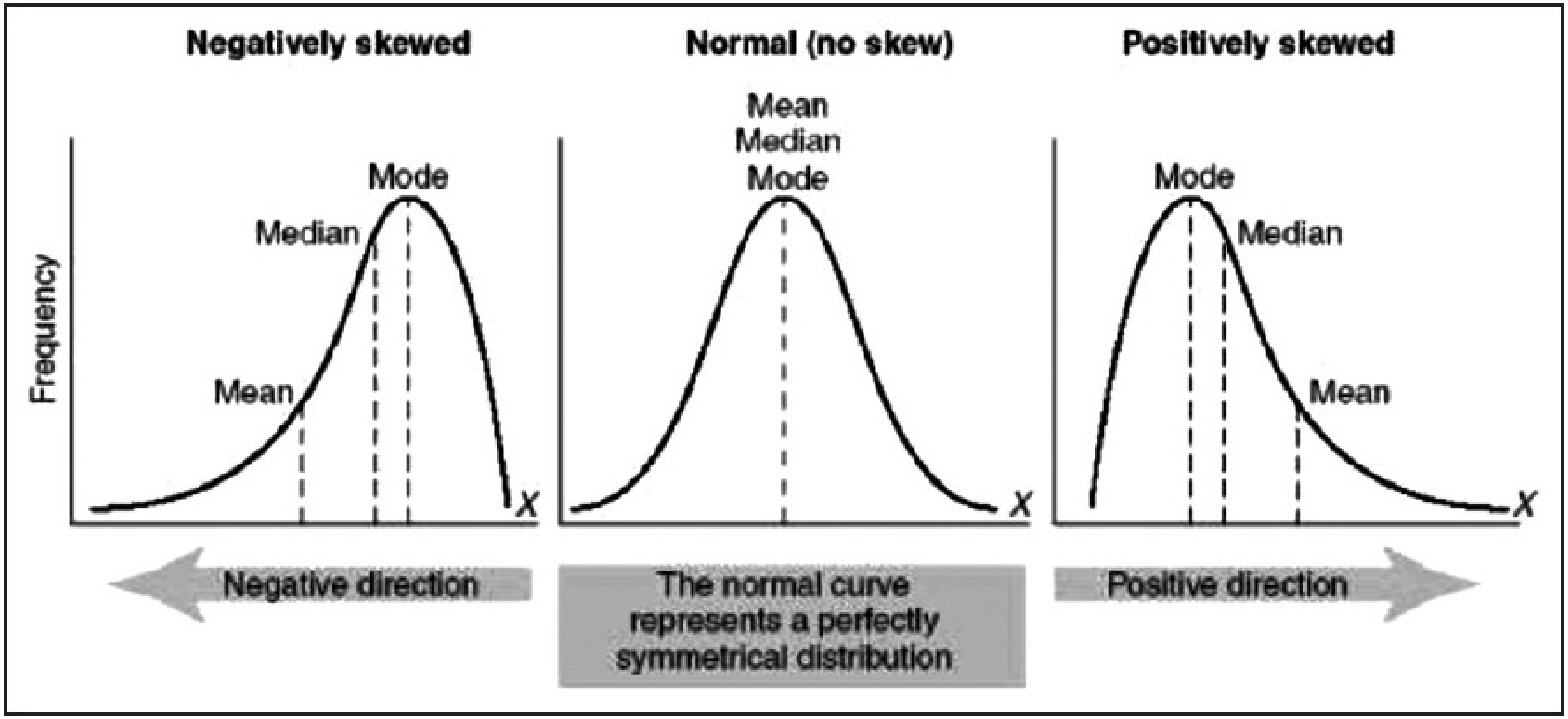
3. For the sample variances with estimated mean:

$$E(\tilde{S}^2) = \frac{n-1}{n}\sigma^2, \qquad E(S^2) = \sigma^2$$

Each equality holds provided the population moments on the right side exist.

5. Sample statistics distributions

To investigate the sample statistics distributions, we make a separation between Gaussian and non-Gaussian populations for rv X.



5.1. Gaussian populations

THEOREM:

Consider an iid **normal** sample $X_1, \dots, X_n$, hence $\forall i : X_i \sim N(\mu, \sigma^2)$, then

1. $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$

2. $\frac{n\,V^2}{\sigma^2} = \frac{\sum(X_i-\mu)^2}{\sigma^2} \;\sim\; \chi^2_n.$

3. $\overline{X}$ and S^2 are independent rvs.

4. $\frac{(n-1)\,S^2}{\sigma^2} = \frac{n\,\tilde{S}^2}{\sigma^2} = \frac{\sum(X_i-\overline{X})^2}{\sigma^2} \;\sim\; \chi^2_{n-1}.$

5. $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \;\sim\; t_{n-1}.$

6. $\frac{S_1^2}{S_2^2}/\frac{\sigma_1^2}{\sigma_2^2} = \frac{\sum\left(X_i - \overline{X}\right)^2/(n_1-1)}{\sum\left(Y_j - \overline{Y}\right)^2/(n_2-1)}/\frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1-1,n_2-1}$

where

$$\overline{X} = \frac{\sum X_i}{n},\ V^2 = \frac{\sum(X_i - \mu)^2}{n},\ \tilde{S}^2 = \frac{\sum(X_i - \bar{X})^2}{n},\ S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

and where the last item concerns the variance ratio S_1^2/S_2^2 of two independent normal samples: $X_1, \ldots, X_{n_1}$ iid from $X \sim N(\mu_1, \sigma_1^2)$, and $Y_1, \ldots, Y_{n_2}$ iid from $Y \sim N(\mu_2, \sigma_2^2)$, with X and Y independent.

Reminder: For the proof of the former theorem, remember that:

1. If $Z \sim N(0,1)$, then $U = Z^2 \sim \chi^2_1$, i.e. χ^2 **distribution** with 1 df. (df: degree of freedom.)

2. If $U_1, \ldots, , U_n$ are independent χ^2 rvs with 1 df, then $V = U_1 + \ldots + U_n \sim \chi^2_n$, i.e. χ^2 **distribution** with n df.

3. If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ and Z and U are independent, then $\frac{Z}{\sqrt{U/n}} \sim t_n$, i.e. **Student's t distribution** with n df.

4. If U and V are independent χ^2 rvs with resp. m and n df, then $W = \frac{U/m}{V/n} \sim F_{m,n}$, i.e. **F distribution** with m and n df.

PROOF:

1. Since $\overline{X}$ is linear combination of independent normal variables, it is normally distributed with $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$ and $Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \sigma^2/n$. Standardizing gives $N(0,1)$.

2. $X_i \sim N(\mu,\sigma^2) \Rightarrow \frac{X_i-\mu}{\sigma} \sim N(0,1) \Rightarrow \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$.

3. This states *the independence of the sample mean and the sample variance in a normal sample*. An equivalent form is: in a normal sample the statistics $\sum_i X_i$ and $\sum_i (X_i - \overline{X})^2$ are independent.

   - $X_i - \overline{X}$ and $\overline{X}$ are independent if covariance is zero (since bivariate normal)

$$\begin{aligned} Cov(X_i - \overline{X}, \overline{X}) &= Cov(X_i, \overline{X}) - Cov(\overline{X}, \overline{X}) \\ &= \frac{1}{n} \sum_{j=1}^{n} Cov(X_i, X_j) - Var(\overline{X}) \\ &= \frac{1}{n} Cov(X_i, X_i) - \frac{\sigma^2}{n} = 0 \end{aligned}$$

   - Hence $(X_i - \overline{X})^2$ and $\overline{X}$ are independent.

   - Hence $\sum_i (X_i - \overline{X})^2$ and $\overline{X}$ are independent.

4. $\frac{\sum(X_i-\mu)^2}{\sigma^2} = \frac{\sum[(X_i-\overline{X})+(\overline{X}-\mu)]^2}{\sigma^2} = \frac{\sum(X_i-\overline{X})^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$.
This is relation of form $W = U + V$ and since U and V are independent:
$M_W(t) = M_U(t)M_V(t)$.
$\Rightarrow M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$, which is the mgf of a rv with χ^2_{n-1} distribution.

5. $\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{S^2/\sigma^2}}$, which is the ratio of $N(0,1)$ rv to the square root of an independent rv with χ^2_{n-1} distribution divided by its degrees of freedom.

6. Follows immediately from [4].

5.2. Non-Gaussian populations

- In some cases, exact sampling distributions are obtained for non-normal populations.

  Examples:

  1. Sample mean from an exponential population

     If $X_1, \ldots, X_n$ is an iid sample from a population $X \sim Exp(\lambda)$ then

     $$2\lambda n\overline{X} \sim \chi^2_{2n}$$

     or $\overline{X} \sim Gamma\left(n, \frac{1}{n\lambda}\right) \sim \frac{1}{2\lambda n}\,\chi^2_{2n}$.

2. Sample mean from a Cauchy distribution

If $X_1, \ldots, X_n$ are iid as $X \sim Cauchy(\mu, \sigma)$ then

$$\overline{X} \sim Cauchy(\mu, \sigma)$$

- **Question:** But what can we say in general when the population does not follow a normal distribution?

- **Answer: Limit theorems**.

Part 2:

Limit Theorems

1. Law of Large numbers (LLN)

THEOREM:

Let $X_1, \ldots, X_n$ be a sequence of independent rvs with $E(X_i) = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$ for $i = 1, \ldots, n$. Let $\overline{X} = \sum_i X_i/n$, then for $n \to \infty$

$$\overline{X} \xrightarrow{P} \mu$$

PROOF:

Since the X_i are independent we know that

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \quad \text{and} \quad Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

The result now follows immediately from Chebyshev's inequality

$$P(|\overline{X} - \mu| > \varepsilon) \leq \frac{Var(\overline{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{as } n \to \infty$$

1.1. Visualization through simulation

Let $X_1, \ldots, X_n$ be independent Bernoulli rvs with $p = 0.5$
$\Rightarrow E(X_i) = 0.5$ and $Var(X_i) = 0.25$.

- **Example:** Tossing fair coins with recording the number of heads

  - ▷ In the first 10 flips, the coin landed headsup 4 times $\Rightarrow \overline{x} = 0.4$
  - ▷ After 30 tosses, the proportion of heads $(\overline{x})$ was 0.567 and after 200 tosses it was 0.502.
  - ▷ After 10000 tosses, he observed 5067 heads $\Rightarrow \overline{x} = 0.5067 \sim E(X_i)$
  - ▷ **Visualization:** $http : //digitalfirst.bfwpub.com/stats\_applet/stats\_applet\_10\_prob.html$

1.2. Application in Gambling: Casino roulette

- If you are playing roulette, each spin is independent from the previous ones.
- The roulette wheel has 18 red pockets out of 37 (European and French). The number of blacks is the same. That said, the probability for both red and black is 48.65%.
- According to the law of large numbers, the more spins are completed, the closer the results of red and black will be to their theoretical probability.

- **Important:** There is no principle that a small number of spins will coincide with the expected value or that one spin will immediately be "balanced" by the others (also known as **Gambers fallacy**)
  - ▷ The **Gambers fallacy** is the irrational belief that prior outcomes in a series of events affect the probability of a future outcome, even though the events in question are independent and identically distributed.

I know the last five scratch off tickets didn't win, which means this next one must be a winner!

2. Central Limit Theorem (CLT)

Reminder: If $X_1, \ldots, X_n$ are iid, we know

- $E[\overline{X}] = E[X_1]$.
- $Var(\overline{X}) = \frac{1}{n} Var(X_1)$.
- Distribution of sample mean of normal distributed variables

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \text{ and independent } \Rightarrow \overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Main question: But what if variables are not generated from normal distribution?

THEOREM:

If $X_1, \ldots, X_n$ are independent and identically distributed rvs with finite second moment $E(X^2)$, and hence with finite mean μ and variance σ^2, then for $n \to \infty$

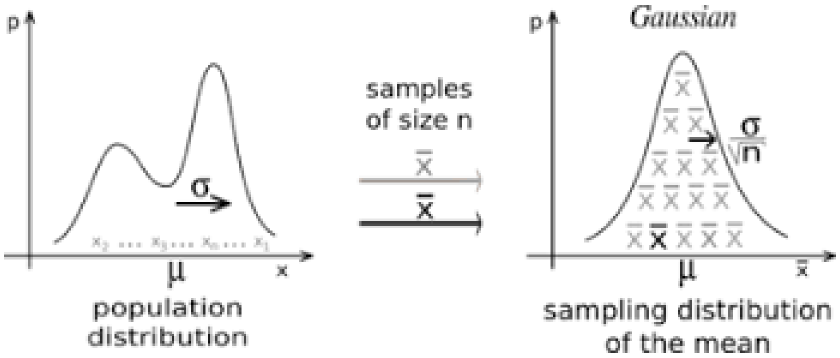
$$P\left[\left(\frac{\overline{X} - E[X_1]}{\sqrt{Var(X_1)/n}}\right) \leq x\right] \to \Phi(x) \qquad \forall x \in \boldsymbol{R}$$

where Φ denotes standard normal cumulative density function.

Said differently, if X_i with $i = 1, \ldots, n$ iid (and n is sufficient large), then

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

with finite $\mu = E[X_1]$ and $\sigma^2 = Var(X_1)$.



Remarks:

- The LLN provides a limiting value for the statistic $\overline{X}$. In particular, if we increase the sample size (if we capture more information from the population), the sample mean converges to the population mean.
- The CLT provides a limiting distribution. Namely $\overline{X}$ is approximately normal with the true mean and the true variance of $\overline{X}$ (provided the number of summands is large).
- In LLN the condition that the X_i have second moments is superfluous (this irrelevant condition in LLN only gives a simple proof).
- If this condition (finite second moments) is added to LLN, the conditions of LLN and CLT are the same.
- When X_i is iid sequence of rvs with finite second moments, LLN and CLT both apply.

- However, then CLT tells us more than LLN, since CLT implies LLN.
- Example: Student distribution with 2 degrees of freedom
  - ▷ Mean exists (equal to 0), but variance does not exist.
  - ▷ LLN holds, but CLT does not.
- Further generalizations weaken the assumption that the X_i have the same distribution and apply to linear combinations of independent random variables. Central limit theorems that weaken the independence assumption and allow the X_i to be dependent (but not too dependent) are also proved.
- It is impossible to give a concise and definitive statement of how good the approximation is for finite values of n, but some general guidelines are available. How fast the approximation becomes good depends on the distribution of the summands (the X_i).

PROOF:

- Consider the standardized mean

$$U_n := \frac{\overline{X} - E[X_1]}{\sqrt{Var(X_1)/n}} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu).$$

- Let $Y_j = X_j - \mu \Rightarrow \alpha_1 = 0$ and $\alpha_2 = \sigma^2$ for Y_j, thus

$$\begin{aligned}\varphi_{Y_j}(t) &= \varphi_{Y_j}(0) + \frac{\varphi^{(1)}_{Y_j}(0)}{1!}t + \frac{\varphi^{(2)}_{Y_j}(0)}{2!}t^2 + o(t^2) \qquad \text{(Taylor)}\\ &= 1 + \frac{i\alpha_1}{1}t + \frac{i^2\alpha_2}{2}t^2 + o(t^2)\\ &= 1 - \frac{\sigma^2 t^2}{2} + o(t^2)\end{aligned}$$

with $\frac{o(t^2)}{t^2} \xrightarrow{t\to 0} 0.$ (negligible rest term)

- Furthermore

$$\begin{aligned}
\varphi_{\frac{X_j-\mu}{\sigma\sqrt{n}}}(t) &= \varphi_{Y_j}\left(\frac{t}{\sigma\sqrt{n}}\right). \\
\Rightarrow \varphi_{U_n}(t) &= \left[\varphi_{Y_j}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n. \qquad \text{(independence)} \\
\Rightarrow \varphi_{U_n}(t) &= \left[1-\frac{t^2}{2n}+o\left(\frac{t^2}{\sigma^2 n}\right)\right]^n \\
&= \left[1+\frac{c_n}{n}\right]^n.
\end{aligned}$$

where $c_n = -\frac{t^2}{2} + \frac{o\left(\frac{t^2}{\sigma^2 n}\right)}{1/n} \xrightarrow{n\to\infty} -\frac{t^2}{2}$

- Hence

$$\varphi_{U_n}(t) \xrightarrow{n\to\infty} e^{-\frac{t^2}{2}} = \varphi_{N(0,1)}(t).$$

Apply continuity theorem of Lévy-Cramér.

2.1. Illustrations

1. Consider an iid sample $X_1, \ldots, X_n$ from an uniform distribution $X \sim U[0,1]$, and consider the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ under increasing sample size n $(n \to \infty)$. Then

   (a) **Limiting value:** $\overline{X} \xrightarrow{P} 1/2$.

   (b) **Limiting distribution:** $\overline{X} \approx N\left(\frac{1}{2}, \frac{1}{12\,n}\right)$.

2. A normal distribution satisfies the LLN and the CLT. For the sample mean of an iid sample $X_1, \ldots, X_n$ from a population $X \sim N(\mu, \sigma^2)$ we know that $\overline{X} \sim N(\mu, \sigma^2/n)$. Hence

   (a) $\overline{X} \xrightarrow{P} \mu$, i.e. the LLN is valid.

   (b) $\overline{X}$ satisfies the CLT-distribution exactly.

Visualizations:

$https://www.zoology.ubc.ca/\sim whitlock/Kingfisher/CLT.htm$

2.2. Normal approximations to different non-Gaussian distributions

- The CLT can also be used to derive normal approximations to several non-Gaussian distributions
- In what follows, we will discuss two well-known distributions, i.e., the χ^2 (continuous) and binomial (discrete) distribution.

1. Normal approximation to a χ^2 distribution

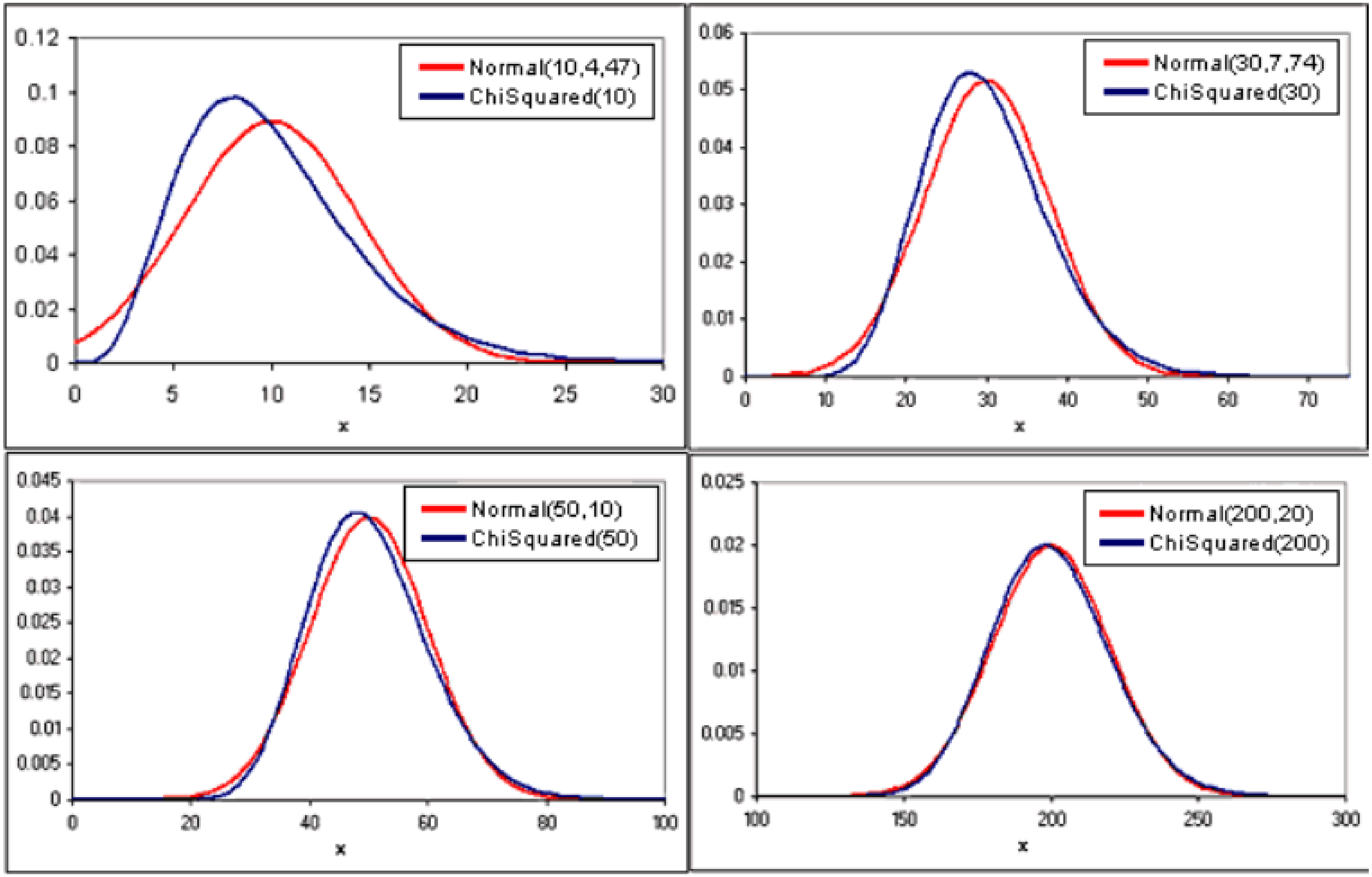
THEOREM:

If n is large, then

$$\chi^2_n \;\approx\; N(n, 2n).$$

PROOF:

Given $X \sim \chi^2_n$, thus $X = Z_1^2 + \ldots + Z_n^2$ with the Z_i iid $N(0, 1)$. If n is large, then $X = Y_1 + \ldots + Y_n,\;\; Y_i = Z_i^2,$ is the sum of a large number of iid variables, and thus approximately normal by the CLT, with its true mean (n) and true variance $(2n)$.


Normal(10,4,47)
ChiSquared(10)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
x
Normal(30,7,74)
ChiSquared(30)
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
x
Normal(50,10)
ChiSquared(50)
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
x
Normal(200,20)
ChiSquared(200)
0.025
0.02
0.015
0.01
0.005
0
100
150
200
250
300
x

2. Normal approximation to a binomial distribution

- ▷ Follows directly from the Theorem of De Moivre - Laplace (1733).
- ▷ Binomial distribution is sum of n independent Bernoulli rvs:

$$Y \sim B(n,p) \Rightarrow Y = X_1 + \ldots + X_n \text{ with } X_i \sim B(1,p).$$

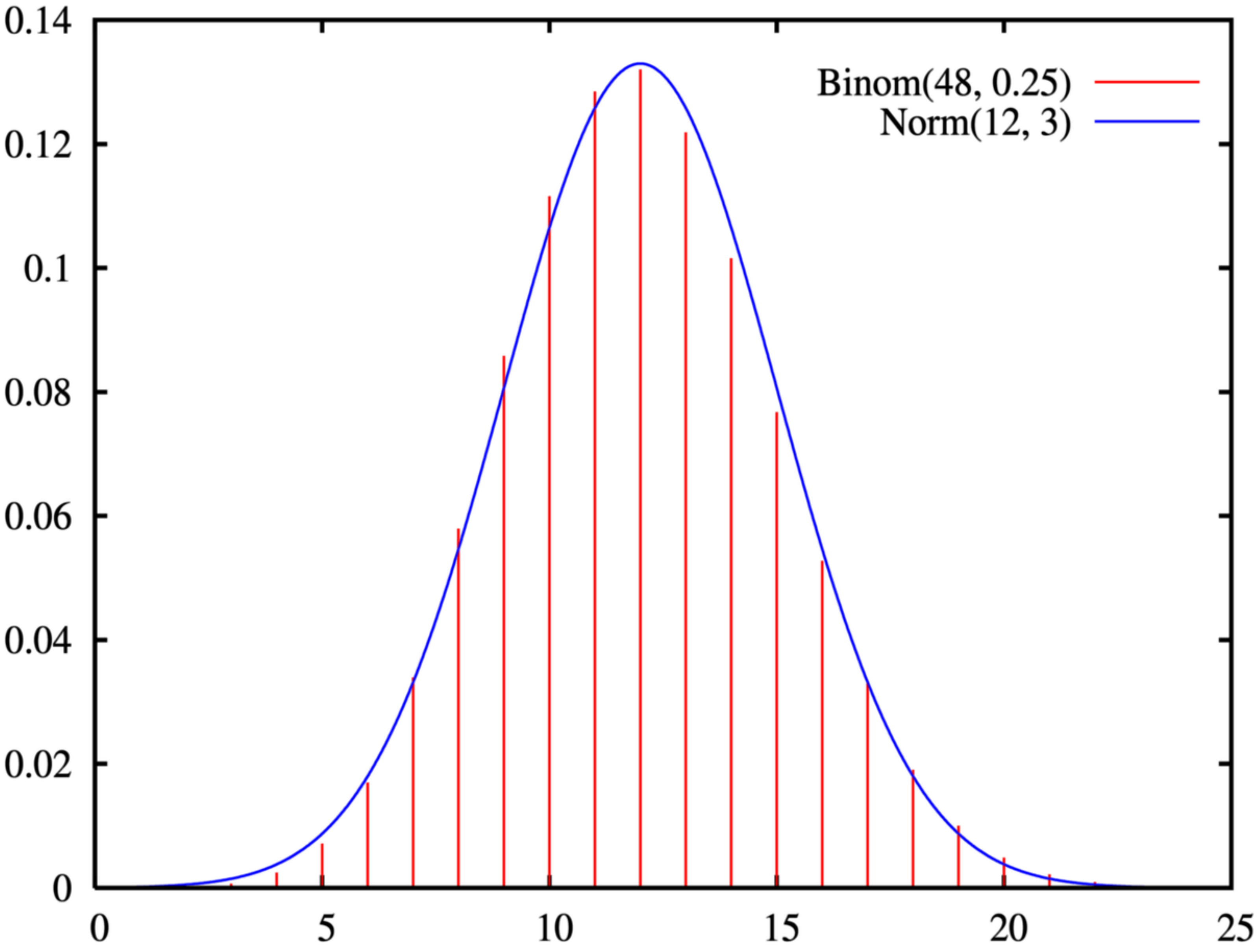
$\Rightarrow E[X_1] = p$ and $Var(X_1) = p(1-p)$.
$\Rightarrow E[Y] = np$ and $Var(Y) = np(1-p)$

- ▷ If n sufficient large,

$$B(n,p) \approx N(np, np(1-p)).$$

Remark:

∗ Approximation is best when $p = 0.5$ (binomial distribution is symmetric) **Rule of thumb**: Approximation is reasonable when $np > 5$ and $n(1-p) > 5$.


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
Binom(48, 0.25)
Norm(12, 3)

2.3. Continuity Correction

Approximating a discrete distribution with a continuous one has shortcomings:

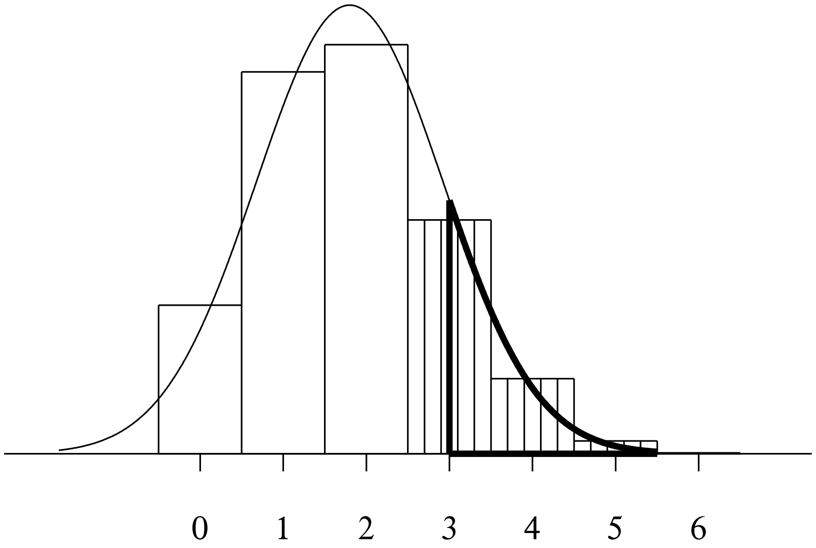
- $P(X \geq a)$ and $P(X > a)$ have different values for a discrete distribution, but will be equal for the continuous approximation.
- Or, $P(X = x) > 0$ for any integer x that is a possible value of X, this probability is necessarily 0 under the approximating pdf.

Example:

- $X \sim B(48, 0.25) \approx Y \sim N(12, 9)$
- $P(X \leq 15) = 0.8768 \approx P(Y \leq 15) = 0.8413$
- Therefore, to approximate a discrete distribution with a continuous distribution $\Rightarrow$ **Continuity Correction**: rounds off integer events to the closest halves.
- $P(X \leq 15) = 0.8768 \approx P(Y \leq 15.5) = 0.8783$ (better approximation)

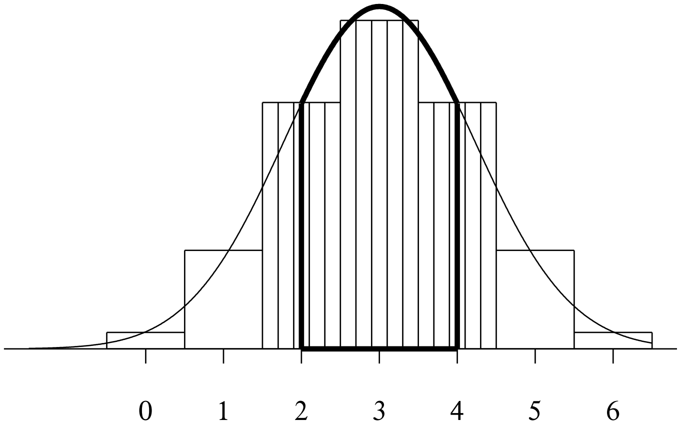
Applying the continuity correction to the normal approximation for the binomial distribution

Reminder: If n sufficient large, $B(n,p) \approx N(np, np(1-p))$.

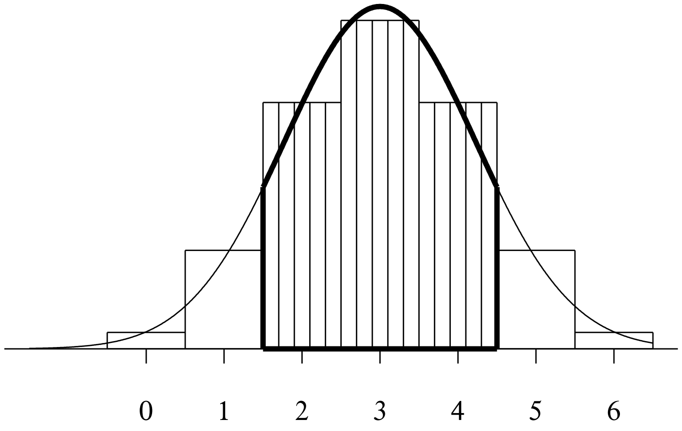


$$P(Y \geq a) \approx 1 - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right) \qquad P(Y \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(a \leq Y \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right)$$



before continuity correction



after continuity correction

2.4. Exercises

Exercise 1:

Suppose a fair coin is tossed 100 times and lands head up 60 times.

1. Write the exact distribution and the approximate normal distribution for the number of heads in case of a fair coin.

2. The expected number of heads is 50. Should we be surprised with the experiment, and doubt that the coin is fair?

Solution exercise 1:

1. Observed rv X: number of heads in 100 tosses.
   - Distribution if coin is fair: $X \sim B(n, p),\ n = 100,\ p = 0.5$
   - Mean and variance of X: $\mu = 50,\ \sigma^2 = np(1-p) = 25 > 5$
   - Normal approximation: $X \approx N(\mu, \sigma^2) = N(50, 25)$.
2. Is the observation $x = 60$ too large compared to the expected value 50? Method of a typical 'statistical test': assume the coin is fair, compute the *significance probability (p-value)* of the observation, $p = P(X \geq 60)$, and doubt fairness of the coin if this probability is too small
   (<5% is a common significance level).

Computation of the prob-value:

- with binomial distribution: $p = P(X \geq 60) = 0.0284$
- normal approx. with continuity correction: $p = P(X \geq 59.5) = 0.0287$
- normal approx. without continuity correction: $p = P(X \geq 60) = 0.0228$

The approximation is improved by the continuity correction.

Test: $p < 0.05$, we should doubt fairness of the coin.

Exercise 2:

Define Y as the number of cities out of 50 with mono-nitrogen oxides $NO_x > 100$ and let $Y \sim \mu g/m^3 \sim B(50, 0.027)$.

Determine $P(Y > 2)$

1. exactly
2. approximated using CLT without continuity correction
3. approximated using CLT with continuity correction

Solution exercise 2:

1.

$$\begin{aligned} P(Y > 2) &= 1 - P(Y \leq 2) \\ &= 1 - \binom{50}{0} 0.027^0 0.973^{50} - \binom{50}{1} 0.027^1 0.973^{49} - \binom{50}{2} 0.027^2 0.973^{48} \\ &= 0.152 \end{aligned}$$

2. $Y \sim N(50 \times 0.027, 50 \times 0.027 \times 0.973)$

$$\begin{aligned} P(Y > 2) &= 1 - P(Y \leq 2) \\ &= 1 - P(Z \leq \frac{2 - 1.35}{\sqrt{1.314}}) \\ &= 1 - P(Z \leq 0.567) = 0.285 \end{aligned}$$

3. $Y \sim N(50 \times 0.027, 50 \times 0.027 \times 0.973)$

$$\begin{aligned} P(Y > 2) &= 1 - P(Y \leq 2) \\ &= 1 - P(Z \leq \frac{2 - 0.5 - 1.35}{\sqrt{1.314}}) \\ &= 1 - P(Z \leq 0.567) = 0.285 \end{aligned}$$

- Note that rule of thumb for approximating a binomial distribution by a normal distribution is actually not satisfied here.
- In this case it might be better to use a **Poisson approximation** (see next slides)

Part 3:

Extra's

1. Poisson approximation for binomial probabilities

THEOREM:

If $X_1, \ldots, X_n$ are iid random variables with $X_i \sim B(n, p_n)$, then for $n \to \infty$ and $p_n \to 0$, such that $np_n \to \alpha$ with $0 < \alpha < \infty$, it holds

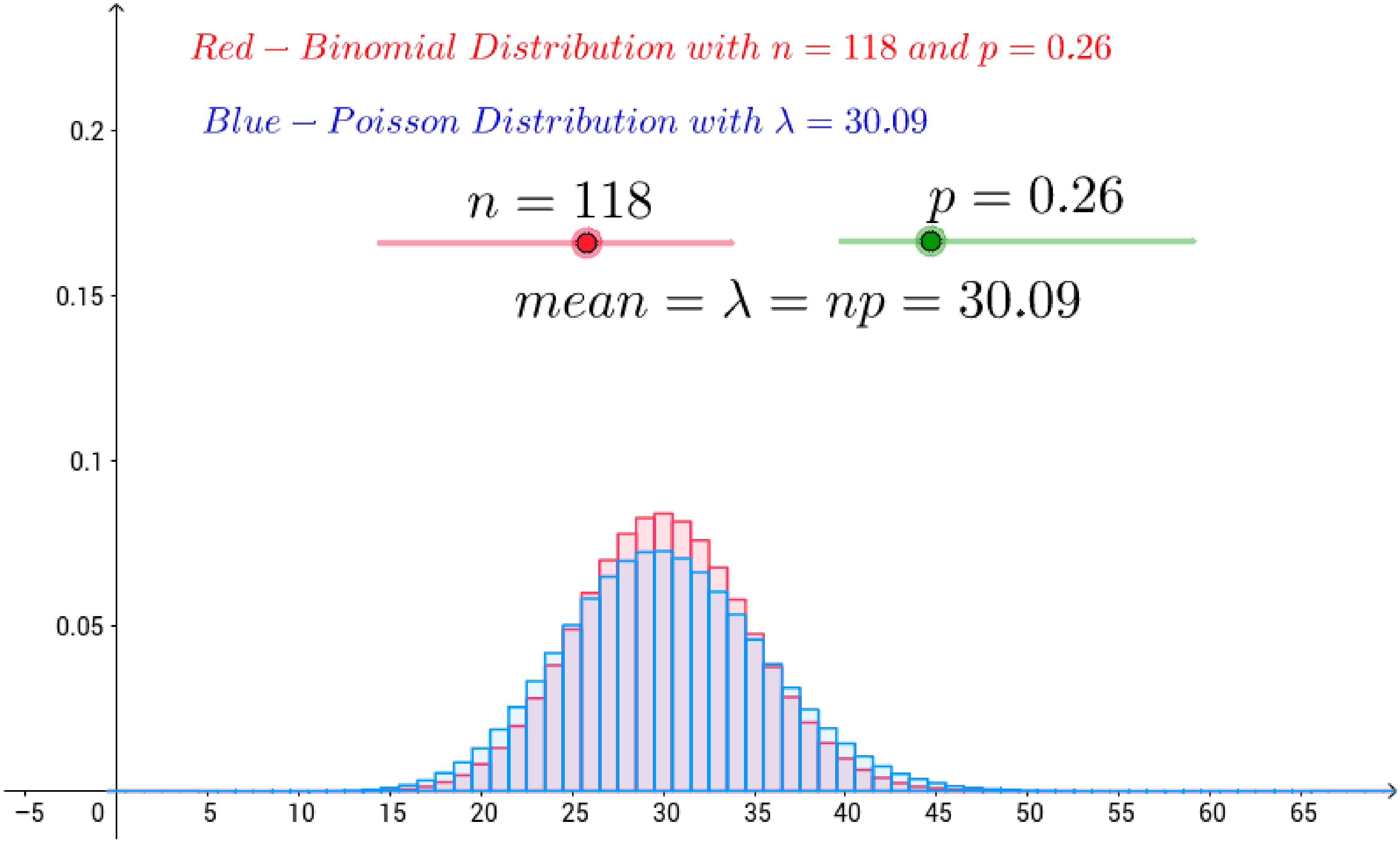
$$X_n \xrightarrow{D} X \qquad \text{with } X \sim Poi(\alpha)$$

Hence

$$B(n, p) \approx Poi(np)$$

Remark:

- $\Rightarrow$ Typically used when p or $1 - p$ is small (and n is large).
  Rule of thumb: $n \geq 30$ and $np < 5$ or $nq < 5$.



Red – Binomial Distribution with $n = 118$ *and* $p = 0.26$

Blue – Poisson Distribution with $\lambda = 30.09$

$n = 118$

$p = 0.26$

$mean = \lambda = np = 30.09$

Binomial distribution

$$P(X = k) = \frac{n!}{k!(n-k!)} p^k (1-p)^{n-k}.$$

Let $np = \lambda$, then we obtain

$$\begin{aligned} P(X = k) &= \frac{n!}{k!(n-k!)} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \end{aligned}$$

If $n \to \infty$

$$\frac{n!}{(n-k)! n^k} \to 1, \quad \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}, \quad \frac{\lambda}{n} \to 0, \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \to 1$$

Hence

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

$$\boxed{\text{PROOF (using mgf):}}$$

$$M_{B(n,p_n)}(t) \xrightarrow{n\to\infty} M_{Ps(\alpha)}(t) \qquad (\forall\, t \in \boldsymbol{R})$$

Computation :

$$\begin{aligned} M_{B(n,p_n)}(t) &= (q_n + p_n e^t)^n \\ &= \left(1 + \frac{(e^t-1)np_n}{n}\right)^n \xrightarrow[n\to\infty]{np_n\to\alpha} e^{\alpha(e^t-1)} \overset{!}{=} M_{Ps(\alpha)}(t) \end{aligned}$$

Using $(1+\frac{a_n}{n})^n \to e^a$ if $a_n \to a$.

2. Normal approximation for Poisson probabilities

THEOREM:

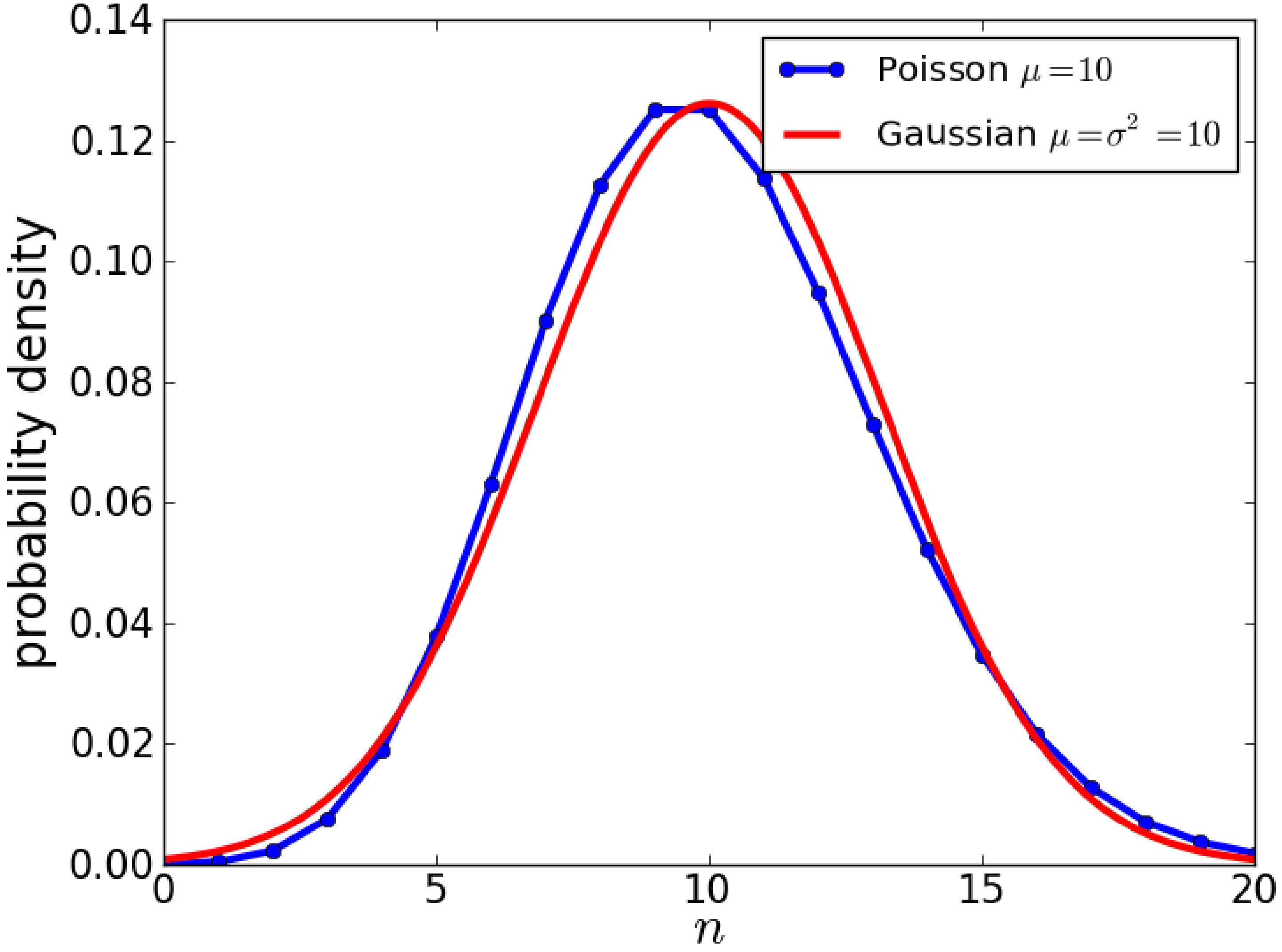
It also holds that

$$Poi(\alpha) \approx N(\alpha, \alpha)$$

Rule of thumb: $\alpha > 10$

Remark:

- This theorem directly follows from the CLT, since the sum of α independent Poisson distributions with parameter 1 follows again a Poisson distribution with parameter α (Proof not shown here).
- Since Poisson is a discrete pdf, the continuity correction can also be applied here.


0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
probability density
0
5
10
15
20
n
Poisson $\mu = 10$
Gaussian $\mu = \sigma^2 = 10$

3. Exercises

Exercise 1:

Let $X \sim Poi(16)$.

Calculate $P(14 \leq X \leq 18)$ exactly and approximated using normal distribution.

Solution exercise 1:

$X \sim Poisson(16) \approx Y \sim N(16,\ 16)$

$p = P(14 \leq X \leq 18) = P(X \leq 18) - P(X < 14) = P(X \leq 18) - P(X \leq 13)$

- Exact (Poisson distribution): $p = 0.4678$
- Normal approximation: $p = 0.3829$
- Normal approximation with continuity correction: $p = 0.4680$

Exercise 2:

Suppose that on average 1/3 of the graduated students of the Bachelors Science bring two extra persons, 1/3 of these students bring one person and 1/3 of these students bring no one extra to the graduation ceremony. If 600 students will be attending the ceremony, what is the probability that not more 650 extra persons will be present?

Solution exercise 2:

- For a random student it holds that X, the number of extra persons at the graduation ceremony, has mean $\mu = 0/3 + 1/3 + 2/3 = 1$ and variance $Var(X) = (0-1)^2/3 + (1-1)^2/3 + (2-1)^2/3 = 2/3$.

- The distribution of W, the total number of extra persons at the graduation ceremony, can be approximated by a normal distribution with mean 600 and variance $600(2/3) = 400$.

- Hence $Z = \frac{W-600}{20} \sim N(0,1)$.

$$P(W \leq 650) = P\left(Z \leq \frac{650.5 - 600}{\sqrt{400}}\right) = P(Z \leq 2.525) = \Phi(2.525) = 0.9942.$$

Exercise 3:

Assume that the number of minutes that are needed to serve a customer at the checkout of a supermarket follows an exponential distribution with parameter $1/3$. Use CLT to calculate the probability that more than 1 hour is needed to serve 16 clients.

Solution exercise 3:

Assume that the number of minutes that are needed to serve a customer at the checkout of a supermarket follows an exponential distribution with parameter 1/3. Use CLT to calculate the probability that more than 1 hour is needed to serve 16 clients.

- Let $X_1, \ldots, X_{16}$ be the times to serve the 16 customers.
- Mean and variance of X_i is 3 and 9.
- Let $Y = \sum_{i=1}^{16} X_i$ be the total time
  $\Rightarrow Y \approx N(16 * 3 = 48, 16 * 9 = 144)$

$$P(Y \geq 60) = 1 - P\left(Z \leq \frac{60 - 48}{\sqrt{144}}\right) = 1 - \Phi(1) = 0.1587.$$

4. Monte-Carlo simulation

DEFINITION:

Monte-Carlo simulation of a parameter θ is an empirical estimation of the parameter, i.e., $\hat{\theta}$, under repeated sampling.

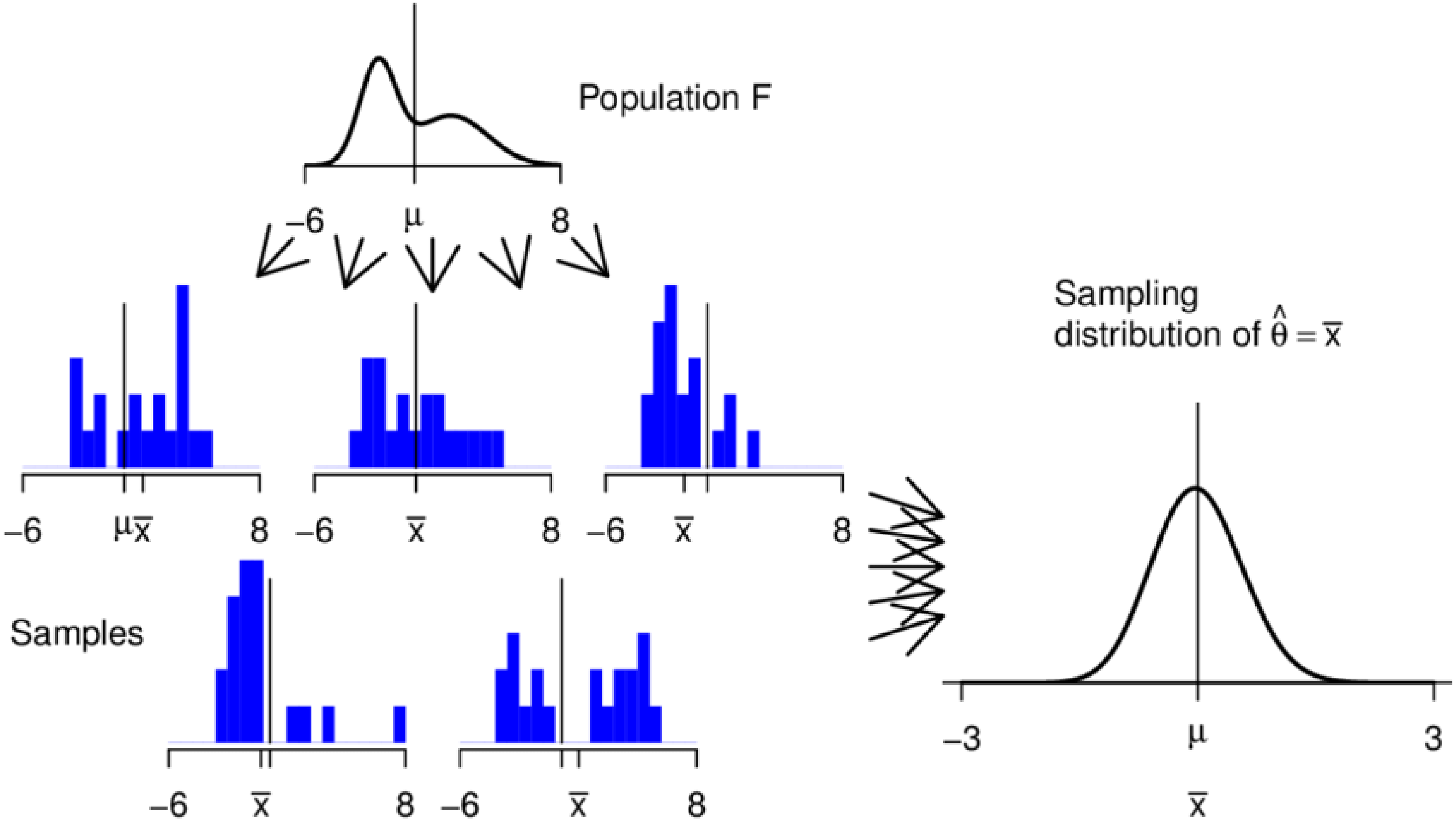
Often,

- the parameter may be seen as a population mean $E(X)$;
- then the simulated value is the corresponding sample mean $\overline{X}$.

Thus,

- the **simulated value** is justified by the LLN,
- the **error on the simulated value** is given by the CLT.

Example: $\theta = \mu_F$



In what follows, we will discuss two illustrations of the Monte-Carlo simulation:

- Monte-Carlo simulation of a probability
- Monte-Carlo simulation of an integral

4.1. Monte-Carlo simulation of an probability

Consider a succes/failure experiment (S/F) with succes probability p.

Estimate p.

- **Method**: Do n independent trials, observe the number of successes X, and estimate p as the sample proportion of successes $\hat{p} = X/n$.
- **Justification**: We give a support for the method, and an estimate for the error in the result.

$p = E(Y)$,where Y_i outcome of random trial, with value 1 for S and 0 for F

$$Y_i \sim B(1, p),\ E(Y_i) = p,\ Var(Y_i) = p(1-p).$$

Observations in n trials: $Y_1, \ldots, Y_n$ are iid.

Simulated value from n trials $\hat{p} = X/n = \overline{Y}$.

Then, for $n \to \infty$:

$$\hat{p} \xrightarrow{D} p, \qquad \hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Conclusion:

- Monte-Carlo simulation for p: $\quad \hat{p} \quad (n$ large$)$
- Error in simulated value: $\quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

  where unknown p is estimated to obtain estimated error $\quad s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

4.2. Monte-Carlo simulation of an integral

Simulate definite integral $I = \int_0^1 g(x)\, dx$ when there is no explicit expression for it.

Generate uniform [0,1] random numbers $X_1, \ldots, X_n$ and compute

$$\hat{I} = \frac{\sum_i g(X_i)}{n} = \overline{g(X)}$$

By the LLN,

$$\hat{I} \xrightarrow{D} E[g(X)] = \int_0^1 g(x)\, f_X(x)\, dx = \int_0^1 g(x)\, dx = I$$

Conclusion:

- Monte-Carlo simulation of I:

$$I = \int_0^1 g(x)\,dx \approx \hat{I} = \frac{\sum g(X_i)}{n}$$

- Error on the estimate:

$$\sigma_{\hat{I}} = \frac{\sigma_{g(X)}}{\sqrt{n}} \approx \frac{|g'(1/2)|}{\sqrt{12\,n}}$$

  where variance on $g(X)$ has been estimated using method of approximate moments, here by linearizing g at $E(X) = 1/2$ (assuming $g'(1/2) \neq 0$).

Example:

Consider the evaluation of

$$I = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2} dx.$$

- From the table of standard normal distribution: I=0.3413.
- Generate 1000 independent uniform rvs $X_1, \ldots, X_{1000}$ on $[0,1]$, then the integral is approximated by

$$\hat{I} = \frac{1}{1000} \left(\frac{1}{\sqrt{2\pi}} \right) \sum_{i=1}^{1000} e^{-X_i^2/2}$$

  which produced for one realization of the X_i the value 0.3417.

Fundamental Concepts of Statistics

Chapter 6: Parameter estimation

Martial Luyts & Clement Cerovecki

Catholic University of Leuven, Belgium

martial.luyts@kuleuven.be

Contents

Part 1:

Statistical estimation problem

1. Introductory material

Suppose we are given the following setting:

- A population rv under investigation: X
- Data sample: $(x_1, \ldots, x_n)$
- Sample: $(X_1, \ldots, X_n)$
- Sample space: set of possible sample values $(x_1, \ldots, x_n)$
(in this course: iid samples, unless stated otherwise).

- The **statistical inference problem**: make (reliable) statements about the distribution of X, based on the data.
- The **parameter estimation problem**: investigate the unknown parameter θ in population X, based on the data.

Reminder: **Parametric estimation**

- ▷ Assume a parametric model for X, i.e.,
  X has density $f(x;\lambda)$ with model parameter λ.
- ▷ We are interesting in investigating parameter $\theta = g(\lambda)$.
  Examples:
  - $*$ $X \sim$ Pois(μ): model parameter μ is investigated parameter $\theta = \mu$,
  - $*$ $X \sim N(\mu, \sigma^2)$: model parameter $\lambda = (\mu, \sigma)$ (parameter vector)
    parameter of interest e.g. $\theta = (\mu, \sigma)$ or $\theta = \mu$ or $\theta = \sigma/\mu$ or $\theta = E(X^2) = \sigma^2 + \mu^2$.

  Often θ is the full or partial model parameter.

To derive statistical inference, three groups of procedures are followed

1. Point estimation

   ⇒ **(point) estimator** for θ: a statistic (a random variable)

   $$T = T\left(X_1, \ldots, X_n\right)$$

   used to estimate or approximate population parameter θ.

   ⇒ **(point) estimate**: numerical value assumed by this statistic when evaluated for a given sample, hence an observed value (a number)

   $$T_{\text{obs}} = T\left(x_1, \ldots, x_n\right)$$

   Example: parameter μ, estimator $\overline{X}$, estimate $\overline{x}$.
   (Note that the estimator is the statistic used to generate the estimate, it is a random variable, whereas an estimate is a number)

2. Confidence intervals (and confidence regions)

3. Statistical tests

- **Task:** Find a good estimator T for θ.

- **Question:** But what are optimal properties of T such that we can say that T is a good estimator for θ?

- **Answer:** Basic philosophy

	Natural requirements for good estimator $T = T\,(X_1, \ldots, X_n)$ for parameter θ	**Formal statistical concepts** on optimality properties of an estimator
1.	T takes values "close" to the true parameter value θ with high probability	• Unbiased estimator • Min variance unbiased estimator • Mean squared error
2.	For large samples, T should be almost perfect estimator: if $n \to \infty$, • T converges to θ • T converges to θ as fast as possible	Asymptotic properties: • Consistent estimator • Asymptotic MSE • Asymptotic relative efficiency
3.	T should absorb all information on θ that is available in the sample	• Sufficient estimator

- In what follows, we will explain these **formal statistical concepts**!

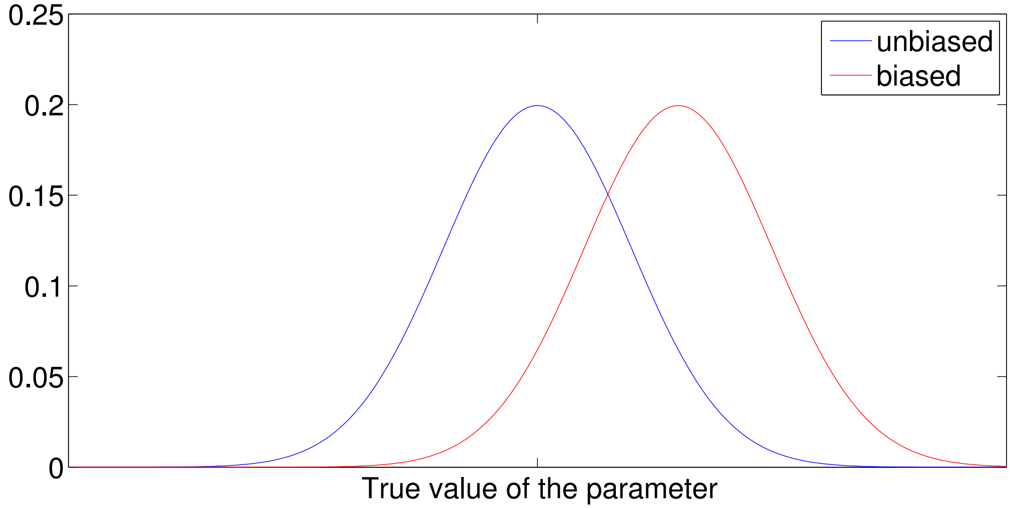
Part 2:

Bias, variance and mean squared error

1. Definitions

Let $T = T\,(X_1, \ldots, X_n)$ be an estimator for θ.

1. - T is an **unbiased** estimator for θ if $\boxed{E(T) = \theta}$

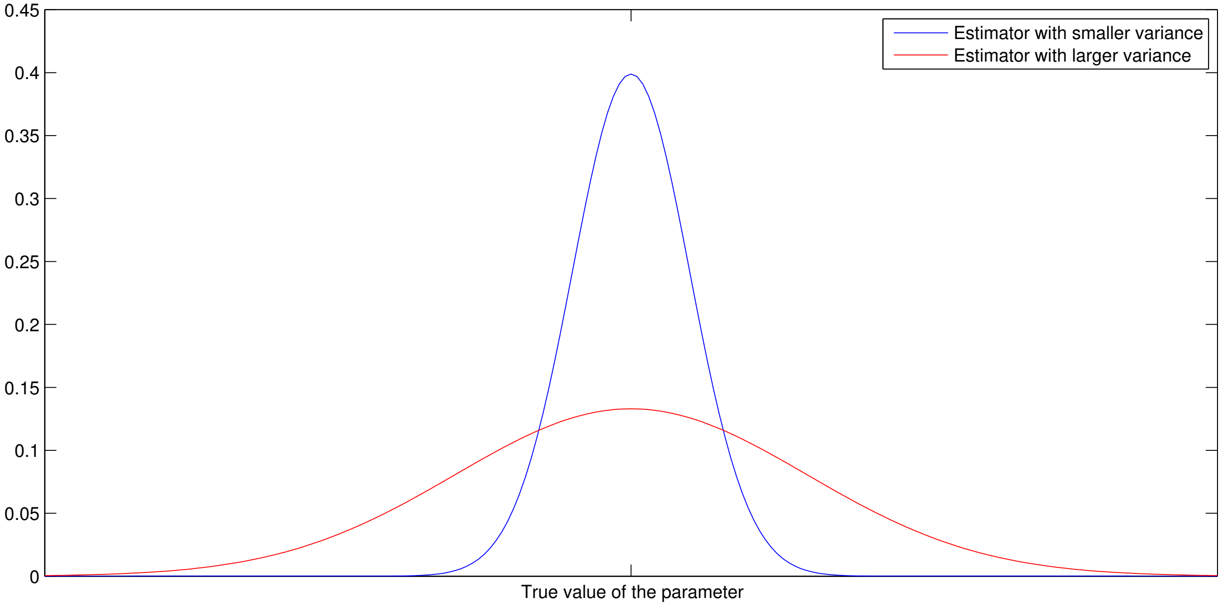


   - The **bias** of estimator T for θ is $\boxed{E(T) - \theta}$

   - T is **asymptotically unbiased** if $\boxed{E(T) \to \theta \text{ as } n \to \infty}$

2. If T and T_0 are two unbiased estimators, then T is **more efficient** than T_0 for θ if

$$\boxed{Var(T) \leq Var(T_0)}.$$



3. The **relative efficiency** of two unbiased estimators T and T_0 for θ is

$$\boxed{\text{eff}\,(T, T_0; \theta) = \frac{Var(T_0)}{Var(T)}}$$

4. The **mean squared error (MSE)** of estimator T for θ is

$$\boxed{\text{MSE}(T;\theta) = E[(T-\theta)^2] = Var(T) + [E(T)-\theta]^2}$$

$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

Thus, the MSE of an unbiased estimator is the variance of the estimator.

5. Let T and T_0 be two estimators for a parameter θ, then T is **more efficient** than T_0 to estimate θ if

$$\boxed{\text{MSE}\,(T;\theta) \leq \text{MSE}\,(T_0;\theta)}$$

2. Exercises on sample statistics

Reminder: Given iid sample $X_1, \ldots, X_n$ from population X.

Parameter	**Corresponding sample statistic**
$\mu = E(X)$	$\overline{X} = \frac{\sum X_i}{n}$
$\sigma^2 = E[(X-\mu)^2]$	$V^2 = \frac{\sum (X_i-\mu)^2}{n}$
	$\tilde{S}^2 = \frac{\sum (X_i-\overline{X})^2}{n}, \ S^2 = \frac{\sum (X_i-\overline{X})^2}{n-1} = \frac{n}{n-1}\tilde{S}^2$
$\alpha_k = E(X^k)$	$a_k = \frac{\sum X_i^k}{n}$
$\mu_k = E[(X-\mu)^k]$	$b_k = \frac{\sum (X_i-\mu)^k}{n}, \quad m_k = \frac{\sum (X_i-\overline{X})^k}{n}$

Exercise 1:

Is $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ an unbiased estimator for σ^2?

Solution exercise 1:

$$\begin{aligned}
E(\tilde{S}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X}\sum_{i=1}^{n}X_i + n\overline{X}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}E(X_i^2) - nE(\overline{X}^2)\right) \\
&= \frac{n}{n}E(X_1^2) - \frac{n}{n}E(\overline{X}^2) \\
&\overset{(*)}{=} (\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2) = \frac{n-1}{n}\sigma^2 \Rightarrow \text{Biased}
\end{aligned}$$

(*) Since $\sigma^2 = E(X_1^2) - \mu^2$, $Var(\overline{X}) = \frac{\sigma^2}{n}$, and $E(\overline{X}) = \mu \Rightarrow E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2$.

Exercise 2:

Let $X_1, \ldots, X_n$ be sample from X, with $X \sim N(\mu, \sigma^2)$. Assume that μ is known and σ^2 is unknown. What is MSE of $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ (for estimating σ^2)?

Solution exercise 2:

V^2 is unbiased estimator since

$$E(V^2) = \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) = \sigma^2$$

Since $X \sim N(\mu, \sigma^2)$, we know that $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_1$.
Therefore

$$\begin{aligned} MSE(V^2) &= Var(V^2) = \frac{1}{n^2}\sum Var\left[(X_1 - \mu)^2\right] \\ &= \frac{1}{n} Var\left[(X_1 - \mu)^2\right] = \frac{\sigma^4}{n} Var\left[\left(\frac{X_1 - \mu}{\sigma}\right)^2\right] = \frac{2\sigma^4}{n} \end{aligned}$$

Exercise 3:

Given the sample variances $\tilde{S}^2$ and S^2 based on an iid sample $X_1, \ldots, X_n$ from a normal population $X \sim N(\mu, \sigma^2)$. Then we know already $T = \frac{(n-1)\, S^2}{\sigma^2} \sim \chi^2_{n-1}$ Find the variance of S^2 and $\tilde{S}^2$.

Solution exercise 3:

$$S^2 = \frac{\sigma^2}{n-1} T \quad \text{where} \quad T \sim \chi^2_{n-1} \Rightarrow Var(T) = 2(n-1).$$

Thus,

$$Var(S^2) = \frac{\sigma^4}{(n-1)^2} Var(T) = \frac{2}{(n-1)} \sigma^4$$

$$Var(\tilde{S}^2) = Var\left(\frac{n-1}{n} S^2\right) = \frac{(n-1)^2}{n^2} Var(S^2) = \frac{2(n-1)}{n^2} \sigma^4$$

3. Best estimator

To find the best estimator T for θ, we prefer an unbiased one with minimum variance, i.e., most efficient of all unbiased estimators.

This estimator is called the **minimum-variance unbiased estimator (MVUE)**.

Definition:

- T is **MVUE (minimum variance unbiased estimator)** for θ if

  (i) $E(T) = \theta$

  (ii) $Var(T) \leq Var(S)$ for any unbiased estimator S for θ

We can restrict this class even further to the **best linear unbiased estimator (BLUE)**.

Definition:

- T is a **BLUE (best linear unbiased estimator)** for θ if

  (i) $E(T) = \theta$ and T is a linear estimator, i.e. of the form

  $$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} c_i X_i$$

  where $c_i \in \boldsymbol{R}$ for $i = 1, \ldots, n$.

  (ii) $Var(T) \leq Var(S)$ for any unbiased and linear estimator S for θ

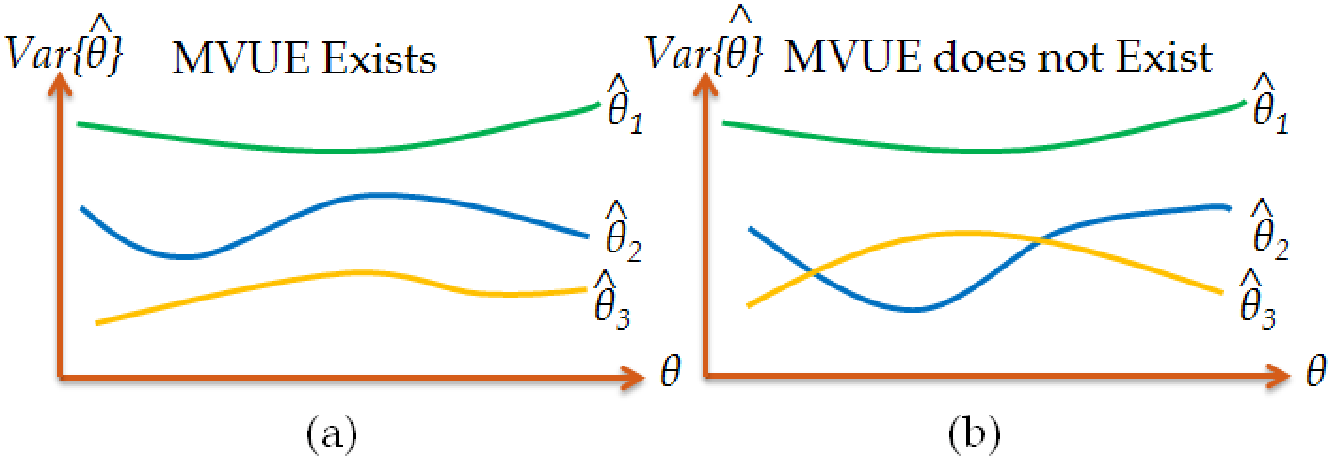
3.1. Existence of minimum-variance unbiased estimator (MVUE)

- Sometimes, there may not exist any MVUE for a given scenario or set of data

- This can happen in two ways:

  1. No existence of unbiased estimators

  2. Even if we have unbiased estimators, e.g., $\hat{\theta}_1, \hat{\theta}_2$ and $\hat{\theta}_3$ for θ, none of them gives uniform minimum variance.



3.2. Methods to find minimum-variance unbiased estimator (MVUE)

There exists methods to find the MVUE:

1. Determine **Cramér-Rao Lower Bound (CRLB)** and check if some unbiased estimator satisfies it.

   If an unbiased estimator exists whose variance equals the CRLB for each value of θ, then it must be the MVUE estimator. It may happen that no estimator exists that achieve CRLB.

   Details about this bound can be found in Section 5.5.

2. Use the **Rao-Blackwell-Lechman-Scheffe (RBLS) theorem**:

   Find a sufficient statistic and find a function of the sufficient statistic.
   This function gives the MVUE.

   - **Remark:** This approach is rarely used in practice.

3. Restrict the solution to find linear estimators that are unbiased. This gives the BLUE.

   - **Remark:** This method gives the MVUE only if the problem is truly linear.

Part 3:

Asymptotic properties

1. Introduction

- When studying the quality of estimators, it is also of interest to look at what happens with the estimators when the sample size n increases and in limit goes to infinity (asymptotic properties).

- In what follows, we will introduce and discuss the terms **consistency** and **asymptotic normality**.

2. Consistency

Denote by $T_n = T(X_1, \ldots, X_n)$ the sequence of rvs $(T_1, T_2, \ldots)$ for which we would like to study its behavior when $n \to \infty$.

DEFINITION:

The sequence of estimators T_n for θ is said to be **consistent** if

$$T_n \xrightarrow{P} \theta \text{ as } n \to \infty$$

This means that

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P\left(|T_n - \theta| > \varepsilon\right) = 0$$

Intuitively, this means that the distribution of the estimator will be more and more concentrated around the true parameter θ as n increases.

- **Question:** Is there a relationship between consistency and unbiasedness? Is every consistent estimator also unbiased and/or is every unbiased estimator also consistent?

- **Answer:** Illustration

  Let $X_1, X_2, \ldots$ be normal iid sample ($X_i \sim N(\mu, \sigma^2)$) and $X \sim N(\mu, \sigma^2)$

  1. Take $T_n = X_1$ as estimator for $E(X)$.
     - $E(T_n) = E(X_1) = \mu$
       $\Rightarrow$ unbiased.
     - But X_1 will not converge in probability to $E(X)$
       $\Rightarrow$ not consistent.

  2. Take $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}$.
     - $E(T_n) = E(\overline{X}_n + \frac{1}{n}) = \mu + \frac{1}{n}$
       $\Rightarrow$ biased.
     - $T_n \xrightarrow{P} E(X) + 0$
       $\Rightarrow$ consistent.

2.1. Consistency on sample statistics

Reminder: Given iid sample $X_1, \ldots, X_n$ from population X.

Parameter	Corresponding sample statistic
$\mu = E(X)$	$\overline{X} = \frac{\sum X_i}{n}$
$\sigma^2 = E[(X-\mu)^2]$	$V^2 = \frac{\sum (X_i - \mu)^2}{n}$
	$\tilde{S}^2 = \frac{\sum (X_i - \overline{X})^2}{n}, \; S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{n}{n-1}\tilde{S}^2$
$\alpha_k = E(X^k)$	$a_k = \frac{\sum X_i^k}{n}$
$\mu_k = E[(X-\mu)^k]$	$b_k = \frac{\sum (X_i - \mu)^k}{n}, \quad m_k = \frac{\sum (X_i - \overline{X})^k}{n}$

THEOREM:

Given iid sample $X_1, \ldots, X_n$ from population X. Then, as $n \to \infty$:

1. $\overline{X} \xrightarrow{P} \mu$
2. $V^2 \xrightarrow{P} \sigma^2$
3. $\tilde{S}^2 \xrightarrow{P} \sigma^2, \qquad S^2 \xrightarrow{P} \sigma^2, \qquad S \xrightarrow{P} \sigma$
4. $a_k \xrightarrow{P} \alpha_k$
5. $m_k \xrightarrow{P} \mu_k$

PROOF:

Rely essentially on basic limit theorems, convergence transformation theorems and moment-expansions.

1. Follows directly from the LLN.

2. Asymptotic value of $V^2 = \sum(X_i - \mu)^2/n = \overline{Y}$
   - Define $Y = (X - \mu)^2, \; Y_i = (X_i - \mu)^2$.
   - Then Y_i are iid sample from Y and we can apply the LLN.
   - LLN: $V^2 = \overline{Y} \xrightarrow{P} E(Y) = \sigma^2$

3. Asymptotic value of $\tilde{S}^2$

- Using SS lemma: $\tilde{S}^2 = V^2 - (\overline{X} - \mu)^2$
- First term on the right converges to σ^2
- Second term converges to zero since $\overline{X} \xrightarrow{P} \mu$
- Then, convergence transformation theorem finishes the proof.

Asymptotic value of S^2

- In $S^2 = \frac{n}{n-1}\tilde{S}^2$, the first factor goes to 1, the second to σ^2, and convergence transformation theorem finishes the proof.

Asymptotic value of S

- Use $S^2 \xrightarrow{P} \sigma^2$ and the transformation theorem with $g(x) = \sqrt{x}$

4. ## Asymptotic value for a_k

- Proofs are similar to those for S^2 (a_k is in structure as $\overline{Y}$)

5. ## Asymptotic value for m_k

- In the lemma m_k is written as a polynomial in the a_j $(j = 1, \dots, k)$, which, by multiple use of convergence transformation theorems, converges to the same polynomial in the α_j , and this polynomial is μ_k .

3. Asymptotic normality

DEFINITION:

An estimator T_n is **asymptotically normal (distributed) (AN)** if there exist a $\sigma > 0$ such that for $n \to \infty$

$$\frac{T_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

where σ^2 is the asymptotic variance of the sequence of rvs $\sqrt{n}T_n$. We also write this as

$$T_n \sim AN(\theta, \frac{\sigma^2}{n})$$

3.1. Transformation of asymptotic normal estimators

- Suppose that we are interested in estimating $g(\theta)$, where T_n be an estimator for θ

- An obvious estimator for $g(\theta)$ is $g(T_n)$.

- **Question:** But what can we say about the asymptotic normality result for the estimator $g(T_n)$ for $g(\theta)$, given that we have the asymptotic normality result for T_n?

- **Answer:** Delta method

THEOREM:

Suppose that T_n is an AN estimator for θ

$$\frac{T_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1).$$

If $g : \boldsymbol{R} \to \boldsymbol{R} : x \to g(x)$ is function, differentiable at $x = \theta$, with $g'(\theta) \neq 0$, then

$$\frac{g(T_n) - g(\theta)}{|g'(\theta)\sigma| \,/ \sqrt{n}} \xrightarrow{D} N(0,1)$$

3.2. Asymptotic normality on sample statistics

Reminder: Given iid sample $X_1, \ldots, X_n$ from population X.

Parameter	**Corresponding sample statistic**
$\mu = E(X)$	$\overline{X} = \frac{\sum X_i}{n}$
$\sigma^2 = E[(X-\mu)^2]$	$V^2 = \frac{\sum (X_i-\mu)^2}{n}$
	$\tilde{S}^2 = \frac{\sum (X_i-\overline{X})^2}{n}$, $S^2 = \frac{\sum (X_i-\overline{X})^2}{n-1} = \frac{n}{n-1}\tilde{S}^2$
$\alpha_k = E(X^k)$	$a_k = \frac{\sum X_i^k}{n}$
$\mu_k = E[(X-\mu)^k]$	$b_k = \frac{\sum (X_i-\mu)^k}{n}$, $\quad m_k = \frac{\sum (X_i-\overline{X})^k}{n}$

THEOREM:

Given iid sample $X_1, \ldots, X_n$ from population X.

1. $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$ or $\overline{X} \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$
2. $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \xrightarrow{D} N(0,1)$
3. $V^2,\ \tilde{S}^2,\ S^2,$ each is $AN\left(\sigma^2, \frac{1}{n}(\mu_4 - \sigma^4)\ \right)$
4. $a_k \sim AN\left(\alpha_k, \frac{1}{n}(\alpha_{2k} - \alpha_k^2)\ \right)$
5. $m_k \sim AN\left(\mu_k, \frac{1}{n}(\mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_2\mu_{k-1}^2)\ \right)$

For each statement the highest order population moment under consideration should exist (i.e. should be finite), e.g. for the AN of $\tilde{S}^2$ the fourth moment α_4 should exist.

PROOF:

1. Follows directly from the CLT.

2. T converges to $N(0,1)$

- $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \;/\; \frac{S}{\sigma} \xrightarrow{D} N(0,1)/1 \sim N(0,1).$

3. Asymptotic normality of $V^2 = \sum (X_i - \mu)^2/n = \overline{Y}$

- Define $Y = (X - \mu)^2,\ Y_i = (X_i - \mu)^2$.

- Then Y_i are iid sample from Y and we can apply the CLT.

- CLT: $V^2 = \overline{Y} \sim AN(E(Y), Var(Y)/n\,)$
  - ▷ $E(Y) = \sigma^2$
  - ▷ $Var(Y) = E(Y^2) - [E(Y)]^2 = \mu_4 - \sigma^4$.

Asymptotic normality of $\tilde{S}^2$

Verificative proof (Constructive proof is based on convergence of random vectors).

- We need to show that $\frac{\sqrt{n}(\tilde{S}^2-\sigma^2)}{\sqrt{\mu_4-\sigma^4}} \xrightarrow{D} N(0,1)$.
- Using SS lemma:

$$\frac{\sqrt{n}(\tilde{S}^2-\sigma^2)}{\sqrt{\mu_4-\sigma^4}} = \frac{\sqrt{n}(V^2-\sigma^2)}{\sqrt{\mu_4-\sigma^4}} - \frac{\sqrt{n}(\overline{X}-\mu)^2}{\sqrt{\mu_4-\sigma^4}}$$

- Since V^2 is AN, the first term on the right converges to $N(0,1)$.
- Hence it is sufficient to show that second term goes to 0, i.e. $\sqrt{n}(\overline{X}-\mu)^2 \xrightarrow{D} 0$, or to show

$$Y := n^{1/4}(\overline{X}-\mu) \xrightarrow{D} 0$$

- Check that $E(Y)=0, Var(Y)=\sigma^2/\sqrt{n} \to 0$ and use Chebyshev's corollary.

Asymptotic normality of S^2

- Note that $S^2 = \frac{n}{n-1}\tilde{S}^2$ where $\tilde{S}^2$ is AN.

- Check general lemma: if $Y_n \sim AN(\mu, \sigma_n^2)$ then so is $\frac{n}{n-1}Y_n = (1 + \frac{1}{n-1})Y_n$, provided $(n-1)\sigma_n \to \infty$.

- Therefore $\frac{n/(n-1)Y_n - \mu}{\sigma_n} = \frac{Y_n-\mu}{\sigma_n} + \frac{1}{(n-1)\sigma_n}Y_n \xrightarrow{D} N(0,1),$ by use of convergence transformation theorem and since $(n-1)\sigma_n \to \infty$.

- Finally the rv $\tilde{S}^2$ satisfies the condition $(n-1)\sigma_n \to \infty$.

4. Asymptotic normality for a_k

- Proofs are similar to those for S^2 (a_k is in structure as $\overline{Y}$).

5. Asymptotic normality for m_k

- The proof for the AN of m_k requires a theory of convergence of random vectors, which is outside the scope of this course.

Part 4:

Sufficient estimator

1. Introduction

- Just like we want our estimators to be consistent and efficient, we also want them to be sufficient.

- To intuitively explain this principle, an example will be used

  Example:

  - Suppose we have iid samples $x = (x_1, \ldots, x_n)$ from a known distribution with unknown parameter θ.

  - Imagine we have two people:
    - Statistician A: Knows the entire sample, gets n quantities: $x = (x_1, \ldots, x_n)$.

    - Statistician B: Knows $T(x_1, \ldots, x_n) = t$, a single number which is a function of the samples. For example, the sum or the maximum of the samples.

- Heuristically, $T(x_1, \ldots, x_n)$ is a sufficient statistic if Statistician B can do just as good a job as Statistician A, given less information.

  - For example, if the samples are from the Bernoulli distribution, knowing $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ (the number of heads) is just as good as knowing all the individual outcomes, since a good estimate would be the number of heads over the number of total trials! Hence, we dont actually care the ORDER of the outcomes, just how many heads occurred!

2. Formal definition

DEFINITION:

Let $X = (X_1, \ldots, X_n)$ be sample from distribution with parameter $\theta \in \Theta$ (possibly more dimensional parameter). A statistic $T = T(X_1, \ldots, X_n)$ is a **sufficient statistic** for θ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ does not depend on θ, i.e.,

$$P(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$$

Our example:

- **Reminder:** Statistician A has all the samples $x_1, \ldots, x_n$ but statistician B only has the single number $t = T(x_1, \ldots, x_n)$.

- The idea is, Statistician B only knows T = t, but since T is sufficient, doesn't need θ to generate new samples $X'_1, \ldots, X'_n$ from the distribution.

- This is because $P(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ and since he/she knows T = t, he/she knows the conditional distribution (can generate samples)!

- Now Statistician B has n iid samples from the distribution, just like Statistician A. So using these samples $X'_1, \ldots, X'_n$, statistician B can do just a good a job as statistician A with samples $X_1, \ldots, X_n$ (on average). So no one is at any disadvantage.

3. Neyman-Fisher Factorization Criterion

- **Problem:** The formal definition is often hard to check in practice.

- It turns out that there is a criterion that helps us determine whether a statistic is sufficient.

- This criterion is called the **Neyman-Fisher Factorization Criterion (NFFC)**!

DEFINITION:

Let $X_1, \ldots, X_n$ be iid rvs with pdf $f_X(x;\theta)$. A statistic $T = T(X_1, \ldots, X_n)$ is sufficient for θ if and only if there exist non-negative functions g and h such that the joint pdf $f_X(X_1, \ldots, X_n; \theta)$ factorizes as follows:

$$f_X(X_1, \ldots, X_n; \theta) = g(X_1, \ldots, X_n) \cdot h(T(X_1, \ldots, X_n); \theta)$$

i.e., the joint pdf can be split into a product of two terms: the first term g can depend on the entire data, but not θ, and the second term h can depend on θ, but only on the data thorugh the sufficient statistic T. In other words, T is the only thing that allows $X_1, \ldots, X_n$ and θ to interact!

- **Remark:** We call the joint pdf $f_X(X_1, \ldots, X_n; \theta)$ also the **likelihood of the data**, often written as $L(\theta \mid X_1, \ldots, X_n)$.
  - Since $X_1, \ldots, X_n$ be iid rvs with pdf $f_X(x; \theta)$ here, we have:

    $L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f_X(x_i; \theta)$

- **Property:** The formal definition on slide 42 & NFFC are equivalent!
  - Proof is not given here.

4. Examples

Exercise 1:

Let $x_1, \ldots, x_n$ be iid random samples from $Unif(0, \theta)$. Show that $T(x_1, \ldots, x_n) = max\{x_1, \ldots, x_n\}$ is a sufficient statistic.

Solution exercise 1:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} I_{\{x_i \leq \theta\}} = \frac{1}{\theta^n} I_{\{x_1, \ldots, x_n \leq \theta\}} = \frac{1}{\theta^n} I_{\{max\{x_1, \ldots, x_n\} \leq \theta\}} = \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}$$

Choose $g(x_1, \ldots, x_n) = 1$ and $h(T(x_1, \ldots, x_n); \theta) = \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}$.

Exercise 2:

Let $x_1, \ldots, x_n$ be iid random samples from $Poi(\theta)$. Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic.

Solution exercise 2:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!} = \frac{e^{-n\theta} \cdot \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = \frac{1}{\prod_{i=1}^n x_i!} \cdot e^{-n\theta} \cdot \theta^{T(x_1,\ldots,x_n)}$$

Choose $g(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$ and $h(T(x_1, \ldots, x_n); \theta) = e^{-n\theta} \cdot \theta^{T(x_1,\ldots,x_n)}$.

Exercise 3:

Let $x_1, \ldots, x_n$ be iid random samples from $Bern(\theta)$. Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ is a sufficient statistic.

Solution exercise 3:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i} \cdot (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i} \cdot (1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{T(x_1,\ldots,x_n)} \cdot (1-\theta)^{n-T(x_1,}$$

Choose $g(x_1, \ldots, x_n) = 1$ and $h(T(x_1, \ldots, x_n); \theta) = \theta^{T(x_1,\ldots,x_n)} \cdot (1-\theta)^{n-T(x_1,\ldots,x_n)}$.

Part 5:

Maximum likelihood estimator

1. Introduction

- In statistics, **maximum likelihood estimation (MLE)** is the most important and widespread method of parameter estimation.
  - **Reason:** MLE generally provides a more efficient estimator for the unknown parameter θ than other estimation techniques.

- Generally, a **maximum likelihood estimator (MLER)** is the value of θ that maximizes the likelihood function $L(\cdot)$.

- Often, it is mathematically easier to maximize the log-likelihood function instead of the likelihood function.
  - **Property:** Since the log function is monotonic increasing, maximizing log-likelihood is equivalent to maximizing likelihood.

- For simple examples, it is possible to find an explicit closed form for the MLER.

- With more complex models, there is no explicit formula and hence one must write program that computes log-likelihood and then use optimization software to maximize this function numerically.

- If data are independent, then the likelihood is the product of its marginal densities.

2. Terminology

Assume rv X with density $f_X(x, \theta)$ with parameter θ.

Given data sample $x = (x_1, \ldots, x_n)$.

1. The likelihood (function) of parameter θ for the observation **x** is

$$L(\theta \mid X_1, \ldots, X_n) = f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

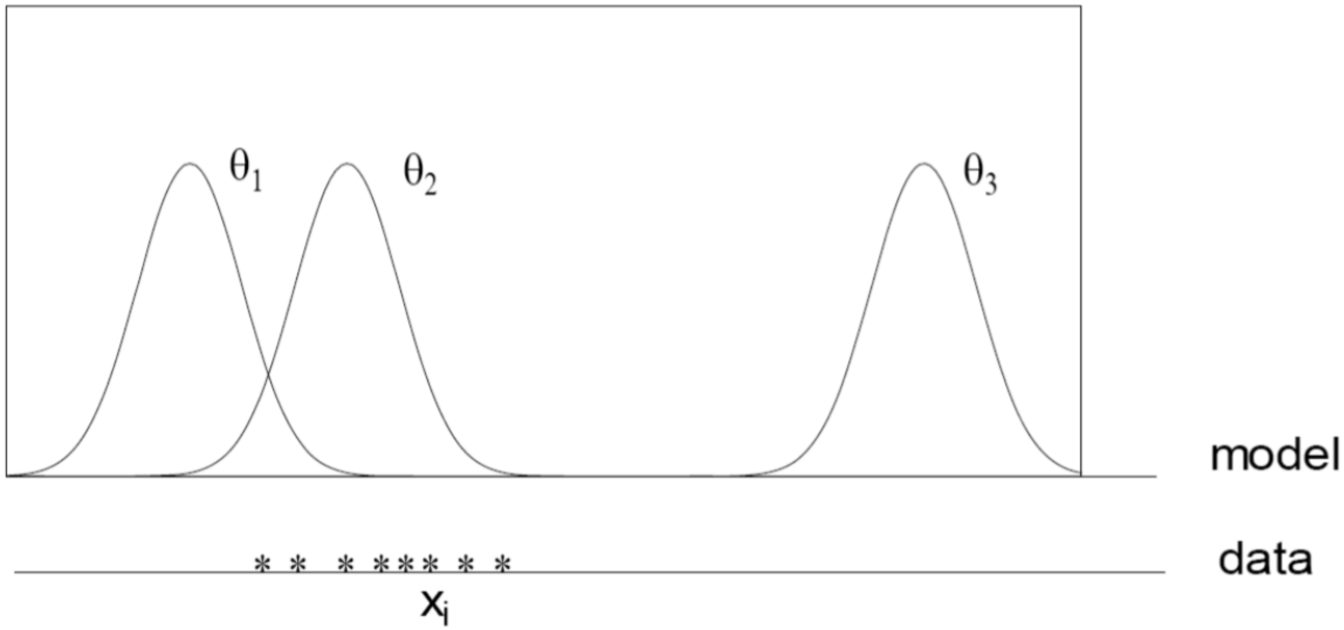
2. The log-likelihood:

$$l(\theta \mid X_1, \ldots, X_n) = log\left[L(\theta \mid X_1, \ldots, X_n)\right] = \sum_{i=1}^{n} log\left[f_X(x_i; \theta)\right]$$

3. The maximum likelihood estimate (MLEr) for θ is the estimate $\hat{\theta}_{ML} = T(x_1, ..., x_n)$. The maximum likelihood estimator (MLER) is then $\hat{\theta}_{ML} = T(X_1, ..., X_n)$.

3. Concept through example

- Given a set of observations $x_1, \ldots, x_n$ of a random sample $X_1, \ldots, X_n$ originating from a parametric model $\{f(x;\theta) \mid \theta \in \underbrace{\{\theta_1, \theta_2, \theta_3\}}_{\Theta}\}$ with unknown parameter θ.



- **Aim:** Estimate the parameter θ, based on the observed data $x_1, \ldots, x_n$.

- To do so, MLE tries to determine the parameter θ for which the observed data have the highest joint probability, i.e., for which the probability of obtaining the observed data is a maximum.

- To examine the joint probability at the observed data sample, MLE makes use of the likelihood function $L(\theta \mid X_1, \ldots, X_n)$

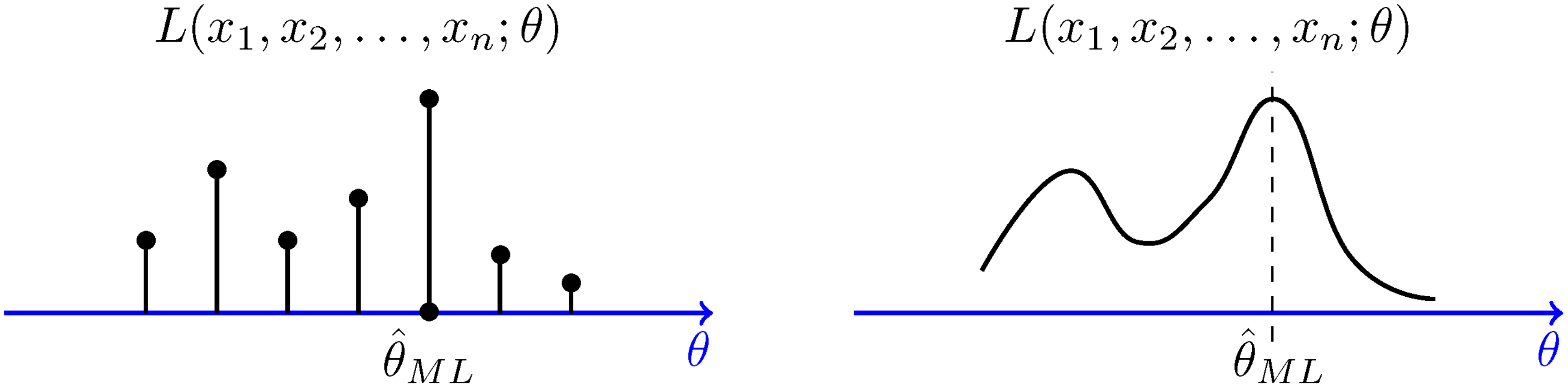
- **Goal of MLE:** Find the value of the model parameter θ that maximize $L(\theta \mid X_1, \ldots, X_n)$ over the parameter space Θ, i.e.,

$$\hat{\theta}_{ML} = \text{argmax}_{\theta \in \Theta = \{\theta_1, \theta_2, \theta_3\}} \{L(\theta \mid X_1, \ldots, X_n)\}$$

  - In the graph above, data are less likely under θ_1, and are absolutely unlikely under θ_3. The highest probability of observing our data here is under θ_2.

  - Thus, $\hat{\theta}_{ML} = \theta_2$

- **Examples:** $\Theta = \mathbb{N}$ (left) & $\Theta = \mathbb{R}$ (right)



- **Procedure for finding the maxima:**
  - Calculate the first derivate of $l(\theta \mid X_1, \ldots, X_n)$ w.r.t. θ, i.e., $D_\theta l(\theta \mid X_1, \ldots, X_n)$.
  - Solve the equation $D_\theta l(\theta \mid X_1, \ldots, X_n) = 0$ to θ.
  - Check for maximum (second derivative < 0).

4. Large sample properties of MLE

Under appropriate smoothness conditions on the pdf

1. The ML estimator from an iid sample is **consistent**:

$$\hat{\theta}_{ML} \xrightarrow{P} \theta.$$

2. The ML estimator is **asymptotically normal**:

$$\hat{\theta}_{ML} \xrightarrow{D} AN\left(\theta, \frac{1}{I_n(\theta)}\right).$$

$I(\theta)$ is called the **Fisher information** about θ in single observation X, given by

$$I(\theta) = E\left\{[D_\theta log f(X;\theta)]^2\right\} = E\left[D_\theta^2 log f(X;\theta)\right] = Var\left[D_\theta log f(X;\theta)\right]$$

Fisher Information $I_n(\theta)$ for sample $(X_1, \ldots, X_n)$ equals n times the information in a single observation: $I_n(\theta) = n \cdot I(\theta)$

3. The multiparameter case: If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a k-dimensional parameter, then the vector of ML estimators is still asymptotically normal:

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{D} AN_k\left(\boldsymbol{\theta}, \frac{1}{n} \cdot \mathbf{I}^{-1}(\boldsymbol{\theta})\right),$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix for a single observation, defined as a $(k$ x $k)$ matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \Big[I_{ij}(\boldsymbol{\theta})\Big]_{i=1,\ldots,k;j=1,\ldots,k},$$

$$I_{ij}(\boldsymbol{\theta}) = E\left\{\left[D_{\theta_i} log f(X;\theta)\right]\left[D_{\theta_j} log f(X;\theta)\right]\right\}$$

5. Fisher information & Cramér-Rao variance bound

- **Question:** Does there exist a lower bound for the variance of an unbiased estimator T for θ?

- **Answer:** Cramér-Rao Lower Bound (CRLB)

THEOREM:

Under suitable regularity conditions (outside the scope of this course), any unbiased estimator T for parameter θ has variance

$$Var(T) \geq \frac{1}{I_n(\theta)}$$

The right side, i.e., $\frac{1}{I_n(\theta)}$, is called the **Cramér-Rao Lower Bound (CRLB)** for an unbiased estimator.

DEFINITION:

An unbiased estimator with variance equal to the CRLB is called an **efficient estimator** (and is the MVUE estimator; Section 3.2).

- **Property:** A ML estimator is an **asymptotically efficient estimator**.

6. Examples

Exercise 1:

Assume that $X_1, \ldots, X_n$ are iid Bernoulli distributed rvs with unknown parameter $\theta = p$. Find the MLER and its asymptotic distribution for θ.

Solution exercise 1:

1. For any observed values $x_1, \ldots, x_n$ (each x_i is 0 or 1), the likelihood and log-likelihood function are given by

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}, \\ \ell(p) &= \sum_{i=1}^{n} \left[x_i \log p + (1-x_i)\log(1-p)\right] \\ &= \log p \sum_{i=1}^{n} x_i + \log(1-p)(n - \sum_{i=1}^{n} x_i). \end{aligned}$$

Hence we get

$$D_p\ell(p) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^n x_i}{n}.$$

This is indeed **maximum** (check!) and thus the MLER of p is $\underline{\hat{p} = \bar{X}}$.

2. Information about p in a single observation $X \sim B(1,p)$:

$$D_p^2\ell(p) = -\left(\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right) \Rightarrow I(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

3. **Asymptotic distribution:** $\underline{\hat{p} \sim AN(p, \frac{p(1-p)}{n})}$.

Exercise 2:

Assume that $X_1, \ldots, X_n$ are iid Poisson distributed rvs with unknown parameter $\theta = \lambda$. Find the MLER and its asymptotic distribution for θ.

Solution exercise 2:

1. The log-likelihood (for observed values $x_1, \ldots, x_n$) is given by

$$\ell(\lambda) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda - \log x_i!) = \log \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log x_i!.$$

Setting the 1st derivative of log-likelihood equal to zero gives us

$$D_\lambda \ell(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Check that 2nd derivative is negative! Thus, the MLER of λ is $\underline{\hat{\lambda} = \bar{X}}$.

2. Information about λ in a single observation:

$$I(\lambda) = E\left[-D_\lambda^2 \log f(X,\theta)\right] = E\left[-\left(-\frac{X}{\lambda^2}\right)\right] = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

3. **Asymptotic distribution:** $\underline{\widehat{\lambda} \sim AN\left(\lambda, \frac{\lambda}{n}\right)}$.

Exercise 3:

Assume normal distributed data

$$f_X(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Determine the CRLB for the unbiased estimator $\hat{\mu}$ for μ.

Solution exercise 3:

$$\begin{aligned} \log f_X(x;\mu) &= -\frac{(x-\mu)^2}{2\sigma^2} - \log(\sigma) - \frac{1}{2}\log(2\pi) \\ \frac{\partial}{\partial\mu}\log f_X(x;\mu) &= \frac{x-\mu}{\sigma^2} \end{aligned}$$

Hence

$$I(\mu) = E\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4}E\left[(X-\mu)^2\right] = \frac{1}{\sigma^2}$$

Therefore, any unbiased estimate $\hat{\mu}$ of μ satisfies

$$E[\hat{\mu}] = \mu \qquad \text{and} \qquad Var[\hat{\mu}] \geq \frac{\sigma^2}{n}.$$

Since $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$, the sample mean $\overline{X}$ is the MVUE estimator for the population mean under the Gaussian population.

Fundamental Concepts of Statistics

Chapter 7: Confidence intervals

Martial Luyts & Clement Cerovecki

Catholic University of Leuven, Belgium

martial.luyts@kuleuven.be

Contents

Part 1:

Confidence Intervals

1. Introductory material

- Until now, we discussed properties and methods for deriving a (good) estimator $\hat{\theta} = T$ for an unknown parameter θ.

- **Question:** Can we add more information to an estimator $\hat{\theta}$?

- **Answer:** Yes, confidence intervals!
  - We can find an interval (T_L, T_U) that we think has high probability of containing θ.

  - The length of such an interval gives us an idea of how closely we can estimate θ.

- **Principle:** From point estimation to set estimation

  1. First, we derive the MLEr for a parameter θ, i.e., the most plausible parameter value after observations x have been made.

  2. Then, confidence intervals are (commonly) used in conjunction with point estimates to convey information about the uncertainty of the estimates.

     - A confidence interval for a population parameter θ is a random interval, calculated from the sample, that contains θ with some specified confidence.

     - **Example:** A 95% confidence interval for θ:
       If we were to take many random samples and calculate confidence interval for each sample, about 95% of these intervals would contain θ.

2. Definition

DEFINITION:

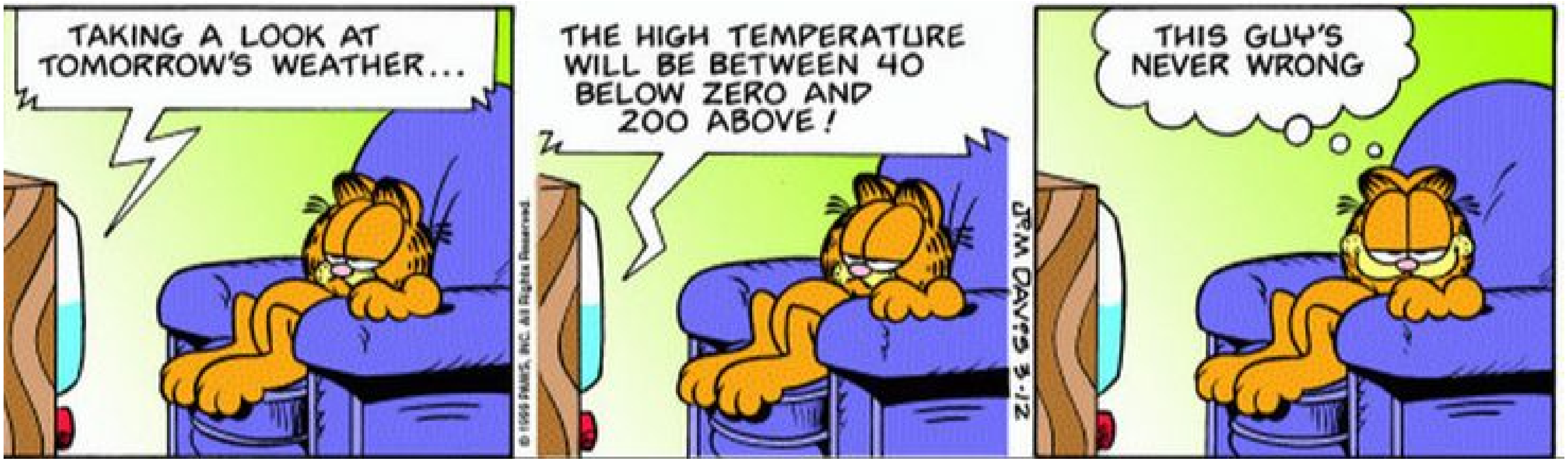
Let **X** $= (X_1, ..., X_n)$ be iid sample from distribution with parameter θ and let α be number with $0 < \alpha < 1$.

1. A $100(1-\alpha)\%$ **confidence interval (CI)** for parameter θ is a random interval $T_L \leq \theta \leq T_U$, where the **confidence limits** $T_L = T_L(\mathbf{X})$ and $T_U = T_U(\mathbf{X})$ are statistics such that the covering probability equals $1-\alpha$, i.e.,

   $$\mathrm{P}\left(T_L \leq \theta \leq T_U\right) = 1 - \alpha.$$

   - T_L and T_U are respectively the **lower confidence limit** and the **upper confidence limit**.

2. The corresponding observed confidence interval based on the data $\mathbf{x} = (x_1, ..., x_n)$ is the numerical interval $[T_L(\mathbf{x}), T_U(\mathbf{x})]$.


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JIM DAVIS 3-12

3. Interpretation

- A choice of $\alpha = 5\%$ gives a 95% confidence interval
- Under long run repeated sampling, about 95% of the observed confidence intervals will cover the true parameter θ, and about 5% will not cover the true parameter θ.
- One way to think of the random interval (T_L, T_U) is to imagine that the sample that we observed is one of many possible samples that we could have observed. Each such sample would allow us to compute an observed interval. Prior to observing the samples, we would expect 95% of the intervals to contain θ. Even if we observed many such intervals, we won't know which ones contain θ and which ones don't.
- **Useful visualisation:** https://rpsychologist.com/d3/ci/

4. Construction procedure

Assume we wish to construct a CI for parameter θ with confidence level $1-\alpha$, given sample $\mathbf{X} = (X_1, ..., X_n)$.

Procedure:

1. Find an estimator for θ, e.g., the ML estimator $\hat{\theta}_{ML}$.
   Obtain an associated **test statistic (pivotal statistic)**: statistic $T(\mathbf{X};\theta)$ with distribution independent of θ.

2. Choose interval $[c_L, c_U]$ for $T(\mathbf{X};\theta)$ with covering probability $1-\alpha$, i.e.,

$$P\left(c_L \le T(\mathbf{X};\theta) \le c_U\right) = 1-\alpha.$$

3. Solve for parameter θ:

$$\mathsf{P}\left(T_L \leq \theta \leq T_U\right) = 1 - \alpha.$$

4. $(1-\alpha)$ confidence interval for θ is

$$T_L \leq \theta \leq T_U.$$

Therefore, the key steps in the construction of a CI are:

1. The choice of statistic T (and/or estimator of θ);

2. The choice of limits c_L and c_U.

In what follows, we show how to construct a $100(1-\alpha)\%$ two-sided CI for:

1. Normal mean under KNOWN variance
2. Normal mean under UNKNOWN variance
3. Normal variance
4. Proportion
5. Difference of 2 independent means $\mu_1 - \mu_2$
6. Difference of 2 dependent means $\mu_1 - \mu_2$
7. Ratio of two variances

4.1 CI for normal mean under known variance

Given iid sample $X_1, \ldots, X_n$ from a normal population $X \sim N(\mu, \sigma^2)$ with known variance.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the mean μ.

Procedure:

1. Estimator for μ: $\overline{X}$

$$\text{CLT:} \qquad \overline{X} \sim N(\mu, \sigma^2_{\overline{X}}), \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

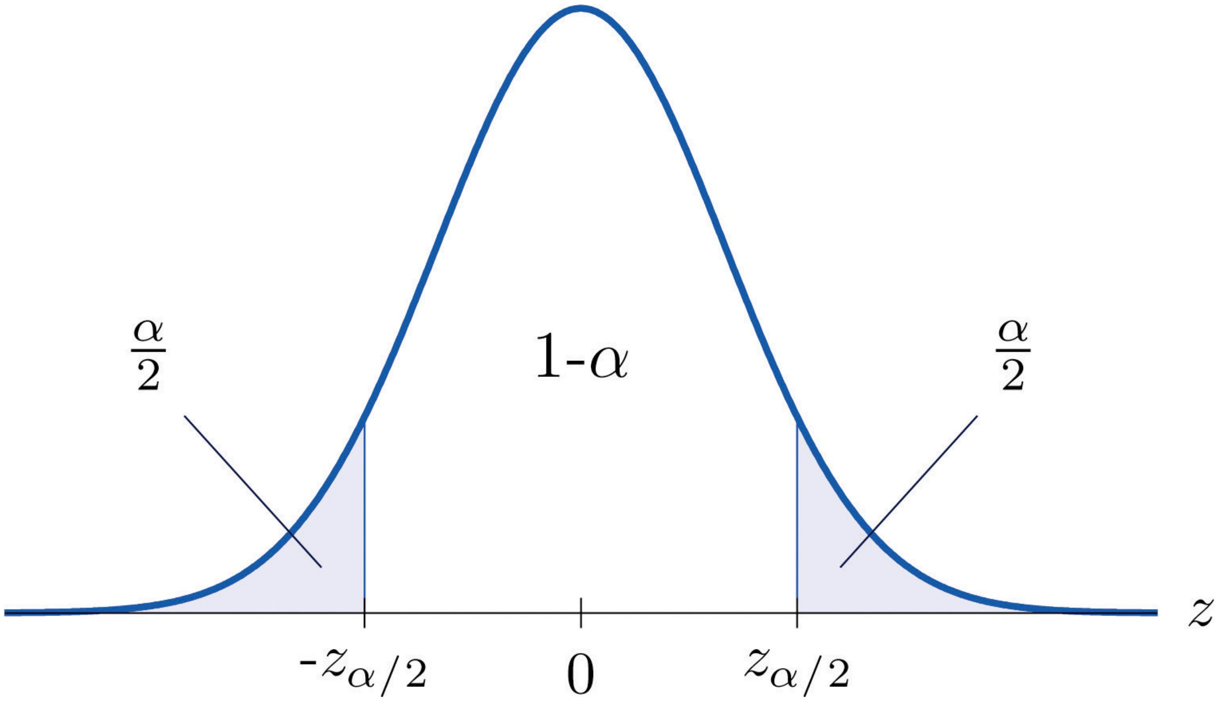
2. Test statistic:

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \sim N(0,1)$$

3. $(1-\alpha)$ probability region: Choose Z-interval symmetric around zero, then

$$P(-z \leq Z \leq z) = 1-\alpha \Leftrightarrow P\left(-z \leq \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \leq z\right) = 1-\alpha,$$

where $z = z_{\alpha/2}$ is $\alpha/2$ upper quantile of standard normal distribution.



4. Solve inequality for parameter μ:

$$P(\overline{X} - z \cdot \sigma_{\overline{X}} \leq \mu \leq \overline{X} + z \cdot \sigma_{\overline{X}}) = 1 - \alpha.$$

5. $(1-\alpha)$ confidence interval for μ is given by

$$\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

   This CI is also written as

$$\left[\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

4.2 CI for normal mean under unknown variance

Given iid sample $X_1, \dots, X_n$ from a normal population $X \sim N(\mu, \sigma^2)$ with unknown variance.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the mean μ.

Procedure:

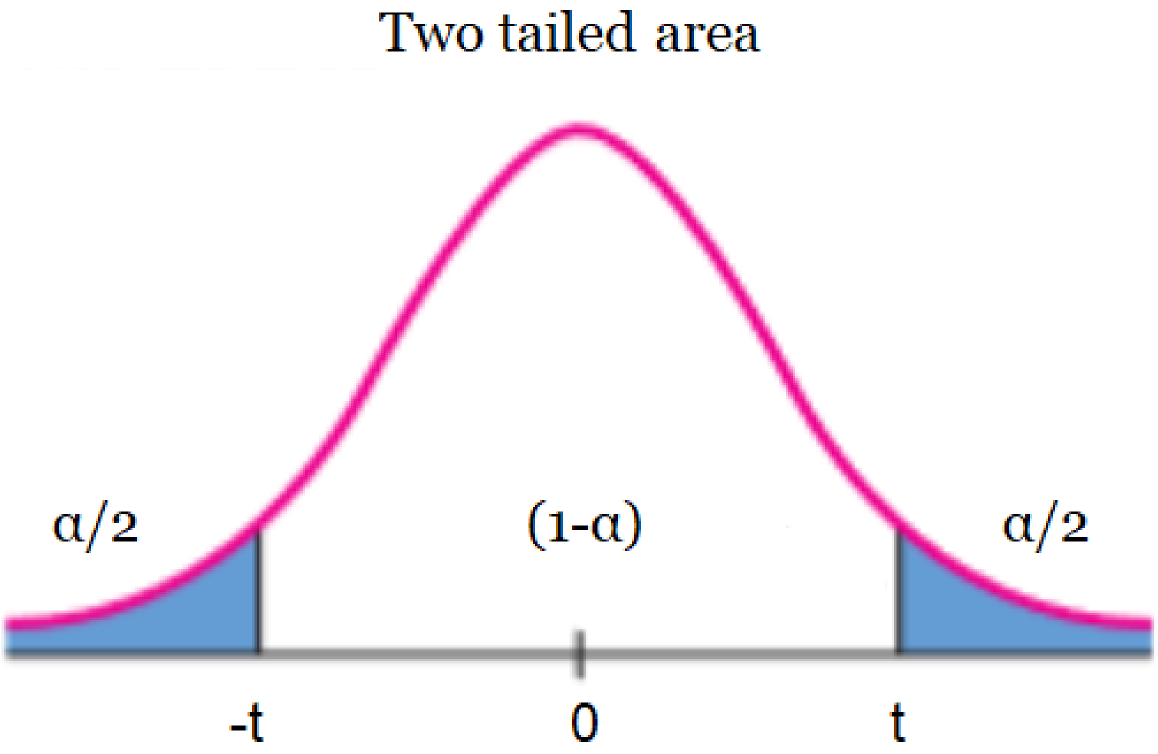
1. Since σ^2 is not known, we will estimate this by the unbiased sample variance $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$
2. Estimator for μ: $\overline{X}$
3. Test statistic:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

4. $(1-\alpha)$ probability region:

$$P(-t \leq T \leq t) = 1-\alpha \qquad P\left(-t \leq \frac{\overline{X}-\mu}{s/\sqrt{n}} \leq t\right) = 1-\alpha$$

where $t = t_{n-1,\alpha/2}$ is $\alpha/2$ upper quantile of t_{n-1}-distribution.



5. Solve inequality for parameter μ:

$$P(\overline{X} - t \cdot s/\sqrt{n} \leq \mu \leq \overline{X} + t \cdot s/\sqrt{n}) = 1 - \alpha.$$

6. $(1 - \alpha)$ confidence interval for μ is given by

$$\overline{X} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

This CI is also written as

$$\left[\overline{X} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$$

4.3 CI for normal variance

Given iid sample $X_1, \ldots, X_n$ from a normal population $X \sim N(\mu, \sigma^2)$ with unknown μ and σ^2.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the variance σ^2.

Procedure:

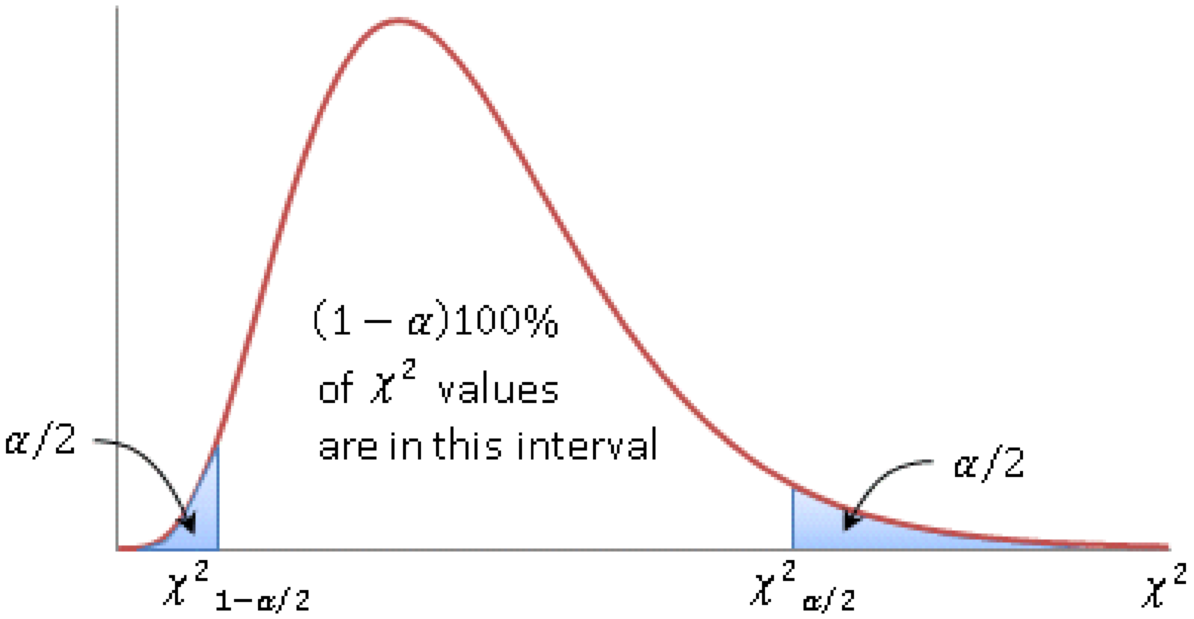
1. Estimator for σ^2:
$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1).$$

2. Test statistic:
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

3. $(1-\alpha)$ probability region:

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

where $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$ are $\alpha/2$ lower and upper quantile of χ^2_{n-1}-distribution.



4. Solve inequality for parameter σ^2:

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha.$$

5. $(1-\alpha)$ confidence interval for σ^2 is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

4.4 CI for proportion

Given iid sample $X_1, \dots, X_n$ from a population $X \sim B(1, p)$ with unknown parameter p.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the proportion p.

Procedure:

1. Estimator for p:
$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$
2. Test statistic:
   - Exact: $n\overline{X} \sim B(n, p)$
   - Approximate (based on CLT): $\overline{X} \approx N(p, \frac{p(1-p)}{n})$

3. $(1-\alpha)$ probability region (based on CLT):

$$\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \Rightarrow P\left(-z \leq \frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z\right) \approx 1-\alpha.$$

4. Since the variance depends on the unknown p, two routes can be followed:

- Approximate variance

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

- Maximize variance

$$\left[\hat{p} - z\sqrt{\frac{1}{4n}}, \hat{p} + z\sqrt{\frac{1}{4n}}\right]$$

Remark: These CI's are based on approximations!

4.5 CI for difference of 2 independent means

SITUATION 1:

Given iid samples of size n_1 and n_2 from normal populations with σ_1 **and** σ_2 **known**.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the mean $\mu_1 - \mu_2$.

Procedure:

1. Estimator for $\mu_1 - \mu_2$: $\overline{X}_1 - \overline{X}_2$

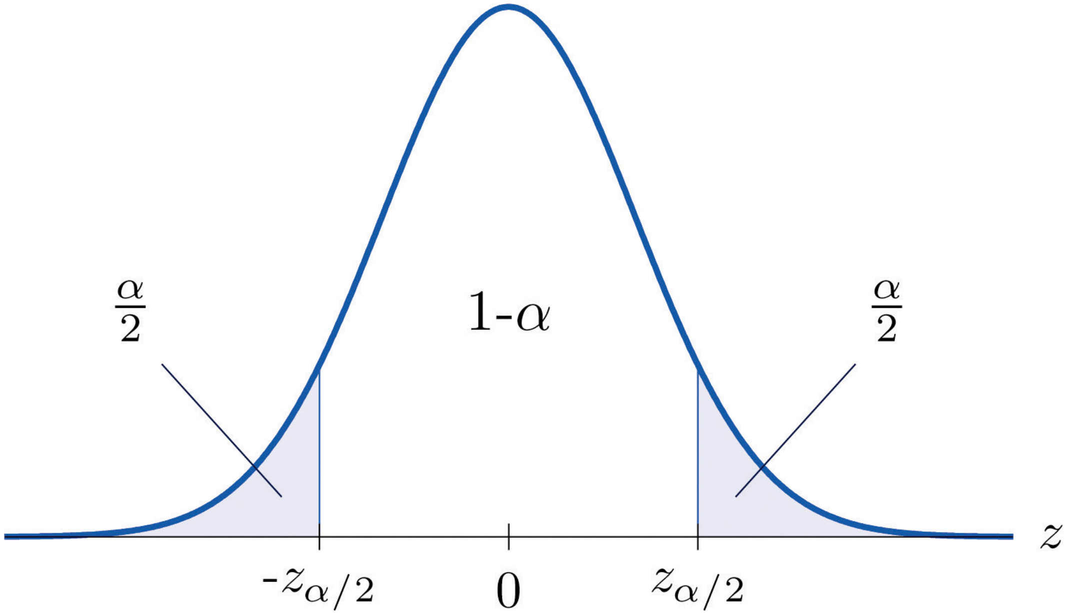
2. Test statistic:

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

3. $(1-\alpha)$ probability region:

$$P\left(-z \leq \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z\right) = 1 - \alpha,$$

where $z = z_{\alpha/2}$ is $\alpha/2$ upper quantile of standard normal distribution.



4. Solve inequality:

$$P\left((\overline{X}_1 - \overline{X}_2) - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{X}_1 - \overline{X}_2) + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

5. $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\left[(\overline{X}_1 - \overline{X}_2) - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; (\overline{X}_1 - \overline{X}_2) + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Special case: When $\sigma_1 = \sigma_2 = \sigma$:

$$\left[(\overline{X}_1 - \overline{X}_2) - z\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\overline{X}_1 - \overline{X}_2) + z\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

SITUATION 2:

Given iid samples of size n_1 and n_2 from normal populations with $\sigma = \sigma_1 = \sigma_2$ **unknown**.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the mean $\mu_1 - \mu_2$.

Procedure:

1. Estimator for $\mu_1 - \mu_2$: $\overline{X}_1 - \overline{X}_2$

2. Test statistic:

We know that

$$\frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}.$$

Since S_1^2 and S_2^2 are independent,

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Problem: σ^2 is not known!

Hence, we make use of a **pooled estimator** S_p^2 for σ^2

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2},$$

which is **unbiased** and based on **all** available observations.

PROOF:

$$E(S_p^2) = \frac{(n_1-1)E(S_1^2) + (n_2-1)E(S_2^2)}{n_1+n_2-2} = \sigma^2 \frac{(n_1-1)+(n_2-1)}{n_1+n_2-2} = \sigma^2.$$

Thus, rewritten, we have that

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

It can be proven that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

PROOF:

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{S_p^2}{\sigma^2}}}$$

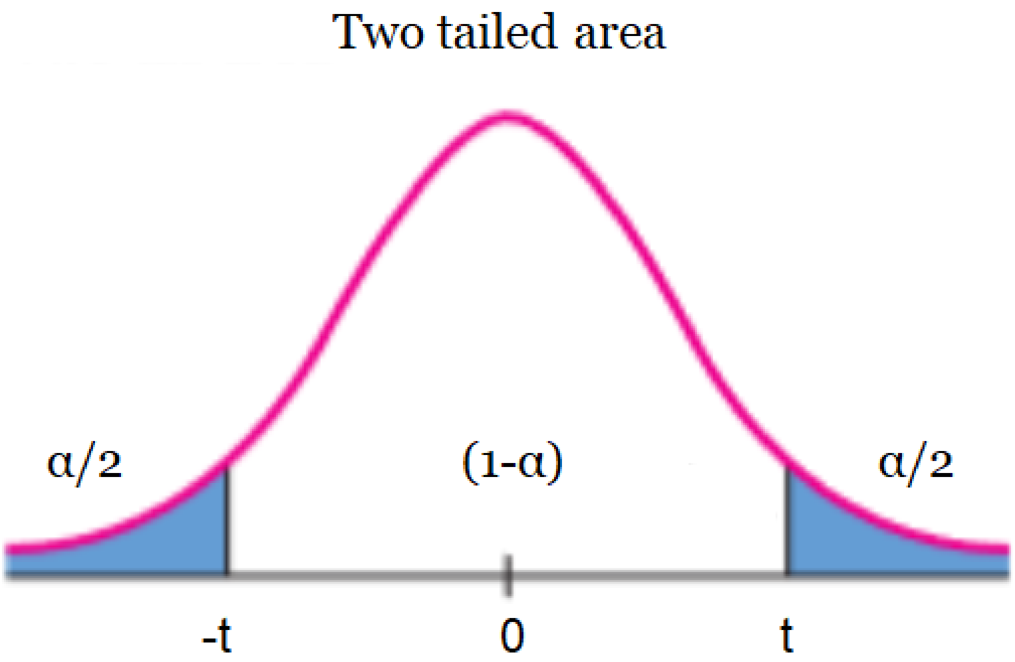
$$= \frac{\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\left[\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}\right]/(n_1+n_2-2)}}$$

$$\sim \frac{N(0,1)}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}} = t_{n_1+n_2-2}$$

3. $(1-\alpha)$ probability region:

$$P\left(-t \leq \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t\right) = 1-\alpha$$

where $t = t_{n_1+n_2-2,\alpha/2}$ is $\alpha/2$ upper quantile of $t_{n_1+n_2-2}$-distribution.



4. Solve inequality for parameter $\mu_1 - \mu_2$, we get the following $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$:

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

SITUATION 3:

Given iid samples of size n_1 and n_2 from normal populations with $\sigma_1 \neq \sigma_2$ **unknown**.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the mean $\mu_1 - \mu_2$.

We call this the **Behrens-Fisher problem**.

Procedure:

1. We replace σ_1 and σ_2 by values of sample standard deviations S_1 and S_2 and proceed as Situation 2 (exercise).

2. A $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$:

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{\nu;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}; (\overline{X}_1 - \overline{X}_2) + t_{\nu;\alpha/2} S_p \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right],$$

where ν is calculated as

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}.$$

4.6 CI for ratio of 2 variances

Given iid samples of size n_1 and n_2 from normal populations with standard deviation σ_1 **and** σ_2, respectively.

Aim: Find a two-sided $100(1-\alpha)\%$ CI for the ratio $\frac{\sigma_1^2}{\sigma_2^2}$.

Procedure:

1. Estimator for $\frac{\sigma_1^2}{\sigma_2^2}$:

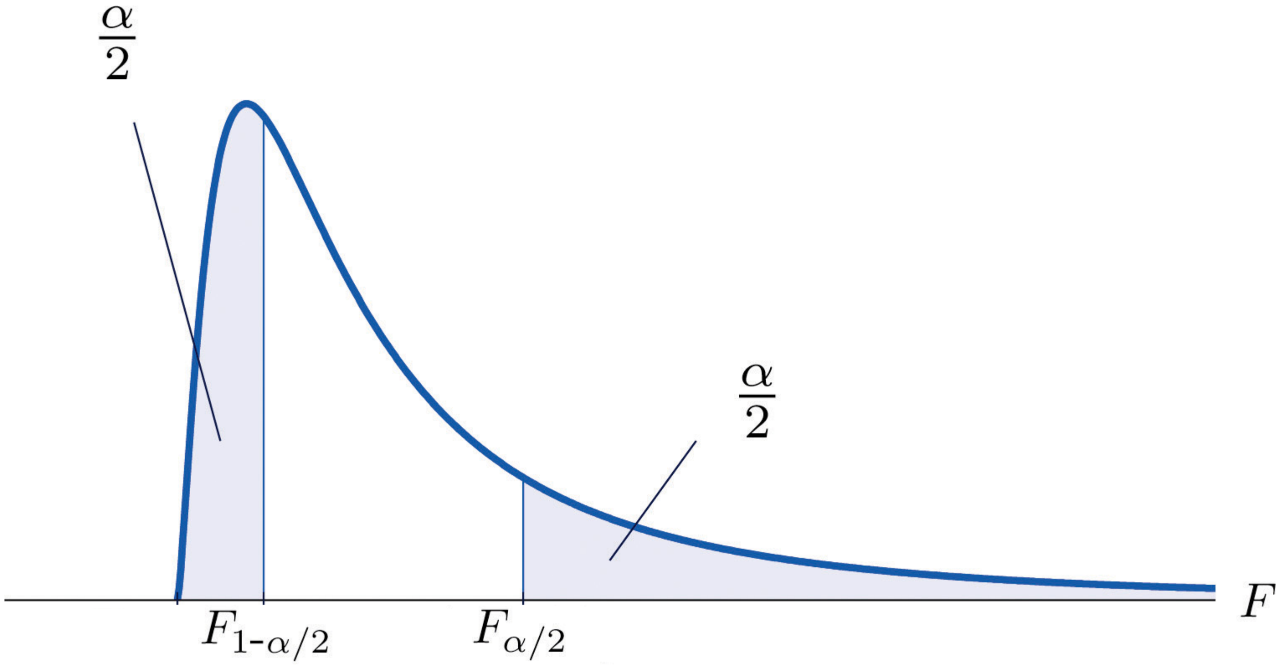
$$\frac{S_1^2}{S_2^2}$$

2. Test statistic:

$$F = \frac{S_1^2}{S_2^2} \Big/ \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

3. $(1-\alpha)$ probability region:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{S_1^2}{S_2^2}/\frac{\sigma_1^2}{\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) \approx 1-\alpha.$$



Remark: $F_{n_1-1,n_2-1,1-\alpha/2} = \dfrac{1}{F_{n_2-1,n_1-1,\alpha/2}}$

4. Solve inequality:

$$P\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{n_1-1,n_2-1,\alpha/2}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{S_1^2}{S_2^2}\cdot F_{n_2-1,n_1-1,\alpha/2}\right)=1-\alpha.$$

5. $(1-\alpha)$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is given by

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{n_1-1,n_2-1,\alpha/2}};\frac{S_1^2}{S_2^2}\cdot F_{n_2-1,n_1-1,\alpha/2}\right]$$

4.7 CI for paired data

- In Section 4.5, we considered CI's for comparing means of two populations based on **independent** samples from each.

- In many experiments, the samples are **paired** rather than independent.
  - ▷ Repeated observations on same sampling unit, e.g. weighting the same individual before and after participating in weight-loss program.
  - ▷ In medical experiment, we might pair subjects who are of same gender and have similar weight and age (e.g. twins) and then one member is randomly assigned to treatment group and the other to control group.

- Pairing can be an effective experimental technique that can control for extraneous sources of variability.
  - ▷ Compare the mean parking speed for two different kind of cars $\Rightarrow$ let same person park both types of cars.

- Because pairing samples makes the observations within each pair dependent, we can not use the methods that were previously developed to compare populations based on independent samples from each.

- The analysis of a matched-pairs experiment uses the n paired differences and inferences regarding the differences in the means are made by making inferences regarding the mean of the differences.

Assume that the differences $D_i = X_i - Y_i$ are a sample from normal distribution (number of observations in group X and Y are equal: $n = n_1 = n_2$), i.e.,

$$\begin{aligned} D_i &\sim N(\mu_D, \sigma_D^2), \\ \mu_D &= E(D_i) = \mu_X - \mu_Y, \\ \sigma_D^2 &= Var(D_i) = \sigma_X^2 + \sigma_Y^2 - 2 \cdot Cov(X,Y). \end{aligned}$$

In general, $Cov(X,Y)$ and hence σ_D is **unknown**.
$\Rightarrow$ inferences will be based on

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}} \sim t_{n-1}.$$

Therefore, a $100(1-\alpha)\%$ CI for μ_D is

$$\left[\overline{D} - t_{n-1,\alpha/2} \cdot s_{\overline{D}}; \overline{D} + t_{n-1,\alpha/2} \cdot s_{\overline{D}}\right].$$

Part 2:

Summary

case		statistic	$(1-\alpha)100\%$ confidence interval
μ			
X normal	σ^2 known	$Z = \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \sim N(0,1)$ $\sigma_{\overline{X}} = \sigma/\sqrt{n}$	$\mu = \overline{X} \pm z_{\alpha/2}\;\sigma_{\overline{X}}$
	σ^2 unknown	$T = \frac{\overline{X}-\mu}{S_{\overline{X}}} \sim t_{n-1}$ $S_{\overline{X}} = S/\sqrt{n}$	$\mu = \overline{X} \pm t_{n-1;\alpha/2}\;S_{\overline{X}}$

$\mu_1 - \mu_2$			
X Y n o r m a l	σ_1^2, σ_2^2 known	$Z = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma_{\overline{X}-\overline{Y}}} \sim N(0,1)$ $\sigma^2_{\overline{X}-\overline{Y}} = \sigma_1^2/n_1 + \sigma_2^2/n_2$	$\mu_1 - \mu_2 = (\overline{X} - \overline{Y}) \pm z_{\alpha/2}\ \sigma_{\overline{X}-\overline{Y}}$
	$\sigma_1^2 = \sigma_2^2$ unknown	$T = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_{\overline{X}-\overline{Y}}} \sim t_{n_1+n_2-2}$ $S^2_{\overline{X}-\overline{Y}} = S_p^2(1/n_1 + 1/n_2)$ $S_p^2 = \frac{\sum(X_i-\overline{X})^2+\sum(Y_i-\overline{Y})^2}{n_1+n_2-2}$	$\mu_1 - \mu_2 = (\overline{X} - \overline{Y}) \pm t_{\nu;\alpha/2}\ S_{\overline{X}-\overline{Y}}$ $\nu = n_1 + n_2 - 2$
	σ_1^2, σ_2^2 unknown, possibly unequal	$T = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_{\overline{X}-\overline{Y}}} \sim t_{\nu}$ $S^2_{\overline{X}-\overline{Y}} = S_1^2/n_1 + S_2^2/n_2$ $\nu \approx \frac{(S_1^2/n_1+S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$	$\mu_1 - \mu_2 = (\overline{X} - \overline{Y}) \pm t_{\nu;\alpha/2}\ S_{\overline{X}-\overline{Y}}$

$\mu_D = \mu_1 - \mu_2$		
Non-independent samples paired obs (X_i, Y_i), $D = X - Y \sim N$	$T = \frac{\overline{D}-\mu_D}{S_{\overline{D}}} \sim t_{n-1}$ $S_{\overline{D}} = S_D/\sqrt{n}$	$\mu_1 - \mu_2 = \overline{D} \pm t_{n-1;\alpha/2}\ S_{\overline{D}}$

p	$X \sim B(1,p), \quad q = 1-p, \quad \hat{q} = 1-\hat{p}$	
$n\hat{p} \geq 5; n\hat{q} \geq 5$ $(np \geq 5; nq \geq 5)$	$Z = \frac{\hat{p}-p}{s_{\hat{p}}} \approx N(0,1)$ $s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$	$p = \hat{p} \pm z_{\alpha/2}\ S_{\hat{p}}$

$p_1 - p_2$	$X \sim b(p_1), \quad Y \sim b(p_2), \quad q_i = 1 - p_i, \quad \hat{q}_i = 1 - \hat{p}_i$	
$n_i\hat{p}_i\hat{q}_i \geq 5$ $(i = 1, 2)$	$Z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{S_{\hat{p}_1-\hat{p}_2}} \approx N(0,1)$ $S^2_{\hat{p}_1-\hat{p}_2} = \hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2$	$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\ S_{\hat{p}_1-\hat{p}_2}$

The construction of the CI for $p_1 - p_2$ is not given in the slides, but it is similar to that for $\mu_1 - \mu_2$ (and it is expected that you are able to do this).

σ^2	X normal	
μ known	$\chi^2 = \frac{nv^2}{\sigma^2} = \frac{\sum(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$	$\frac{nv^2}{c_R} \leq \sigma^2 \leq \frac{nv^2}{c_L}$ $c_R = \chi^2_{n;\alpha/2}, \quad c_L = \chi^2_{n;1-\alpha/2}$
μ unknown	$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$	$\frac{(n-1)S^2}{c_R} \leq \sigma^2 \leq \frac{(n-1)S^2}{c_L}$ $c_R = \chi^2_{n-1;\alpha/2}, \quad c_L = \chi^2_{n-1;1-\alpha/2}$

σ_1^2/σ_2^2	X, Y normal	
μ_1, μ_2 known	$F = \frac{v_1^2/v_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1,n_2}$	$\frac{1}{f_R}\frac{v_1^2}{v_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{f_L}\frac{v_1^2}{v_2^2}$ $f_R = F_{n_1,n_2;\alpha/2}, \quad f_L = F_{n_1,n_2;1-\alpha/2}$
μ_1, μ_2 unknown	$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$	$\frac{1}{f_R}\frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{f_L}\frac{S_1^2}{S_2^2}$ $f_R = F_{n_1-1,n_2-1;\alpha/2}, \quad f_L = F_{n_1-1,n_2-1;1-\alpha/2}$

Part 3:

Examples

1. Exercises

Exercise 1:

Assume a random sample of size $n = 20$ from a normal population with variance $\sigma^2 = 225$ and mean $\overline{X} = 64.3$. Construct a two-sided 95% CI for population mean μ.

Solution exercise 1:

1. Since $z_{0.25} = 1.96$:

$$\left[64.3 - 1.96 \cdot \frac{15}{\sqrt{20}}, 64.3 + 1.96 \cdot \frac{15}{\sqrt{20}}\right]$$

   which reduces to

$$[57.7, 70.9]$$

Exercise 2:

Paint manufacturer wants to determine the average drying time of a new interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and standard deviation of 8.4 minutes, construct a 95% CI for the true mean.

Solution exercise 2:

1. $t_{0.025,11} = 2.201$ gives us

$$\left[66.3 - 2.201 \cdot \frac{8.4}{\sqrt{12}}, 66.3 + 2.201 \cdot \frac{8.4}{\sqrt{12}}\right] \text{ or } [61.0, 71.6]$$

Exercise 3:

In 16 test runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 gallons. Construct 99% CI for σ^2, measuring the true variablility of the gasoline consumption of this engine. Assume that the gasoline consumption of this engine follows a normal distribution.

Solution exercise 3:

1. $\chi^2_{0.005,15} = 32.801$ and $\chi^2_{0.995,15} = 4.601$ gives us a CI for the σ^2:
$$\left[\frac{15(2.2)^2}{32.801}, \frac{15(2.2)^2}{4.601}\right]$$
$\Rightarrow$ CI for σ is $[1.49, 3.97]$ (because of monotonic transformation).

Exercise 4:

A doctor takes a sample of 400 students and it is found that 140 of them are smokers. Find a 95% CI for the proportion of smoking students.

Solution exercise 4:

1. Substituting $\hat{p} = \frac{140}{400} = 0.35$ and $z = 1.96$ leads to the following (approximated) CI for p:

$$\left[0.35 \pm 1.96\sqrt{\frac{(0.35)(0.65)}{400}}\right] \qquad \text{or} \qquad [0.30, 0.40]$$

Exercise 5:

Construct a 95% CI for the actual difference between the average lifetimes of 2 kinds of light bulbs, given that a random sample of 40 light bulbs of one kind lasted on the average 418 hours of continuous use and 50 light bulbs of another kind lasted on the average 402 hours. Population standard deviations are known to be $\sigma_1 = 26$ and $\sigma_2 = 22$.

Solution exercise 5:

1. Since z=1.96, the 95% CI for $\mu_1 - \mu_2$ is

$$\left[(418-402) - 1.96 \cdot \sqrt{\frac{26^2}{40} + \frac{22^2}{50}}, (418-402) + 1.96 \cdot \sqrt{\frac{26^2}{40} + \frac{22^2}{50}}\right]$$

$$= [5.90, 26.10]$$

Exercise 6:

To measure the effect of a certain type of diet 10 overweight volunteers were enlisted for a small scale pilot study. Their weight was measured before and after six months of dieting. The researcher gives you the following results.

n_i	No diet	Diet
1	85	82
2	87	84
3	88	89
4	90	84
5	92	88
6	87	87
7	85	86
8	89	84
9	93	89
10	84	82

Construct a 95% CI for the weight difference before and after six months of dieting.

Solution exercise 6:

n_i	No diet	Diet	Difference
1	85	82	3
2	87	84	3
3	88	89	-1
4	90	84	6
5	92	88	4
6	87	87	0
7	85	86	-1
8	89	84	5
9	93	89	4
10	84	82	2

We find $\overline{D} = 2.5$, $S_D = 2.460804$, hence

$$\left[\overline{D} \pm t_{9,0.975}\frac{S_D}{\sqrt{10}}\right] = \left[2.5 \pm 2.262 * \frac{2.460804}{\sqrt{10}}\right] = [0.7398, 4.2602]$$

Exercise 7:

You are asked to compare the amount of wine in bottles of two different manufactures. You sample 5 bottles of wine of each type, where we assume that the amount of wine of each type comes from normal populations. Construct a 95% CI for the actual difference between the average amount of wine of each type.

n_i	A	B
1	755	756
2	753	755
3	754	754
4	752	754
5	755	756

Solution exercise 7:

1. For type A, you find $\overline{x}_A = 753.8$ ml and for type B you find $\overline{x}_B = 755$. The standard deviations of both samples are respectively 1.30384 and 1.

2. First, we check whether $\underline{\sigma_A = \sigma_B}$.

$$F_0 = \frac{S_A^2}{S_B^2} \sim f(n_A - 1, n_B - 1), \qquad f_0 = \frac{1.30384^2}{1^2} = 1.69999$$

$$f_{4,4,0.025} = \frac{1}{f_{4,4,0.975}} = 0.10411, \qquad f_{4,4,0.975} = 9.60453$$

$$CI : \left[\frac{S_A^2}{S_B^2} f_{4,4,0.025}, \frac{S_A^2}{S_B^2} f_{4,4,0.975}\right] = [0.176986, 16.3276]$$

Since 1 is inside this CI, we assume that $\sigma_A = \sigma_B$.

3. Thus, a 95% CI for the actual difference between the average amount of wine of each type is given by

$$\left[755.3 - 752.75 - 2.021 \cdot S_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}, 755.3 - 752.75 + 2.021 \cdot S_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right]$$

$$= \left[2.55 \pm 2.021 \cdot \sqrt{\frac{(n_A-1)\cdot S_1^2 + (n_B-1)\cdot S_2^2}{n_A+n_B-2}} \cdot \sqrt{\frac{1}{5}+\frac{1}{5}}\right]$$

$$= [1.169553, 3.930447]$$

Fundamental Concepts of Statistics

Chapter 8: Hypothesis testing

Martial Luyts & Clement Cerovecki

Catholic University of Leuven, Belgium

martial.luyts@kuleuven.be

KU LEUVEN

LEUVEN STATISTICS
RESEARCH CENTRE

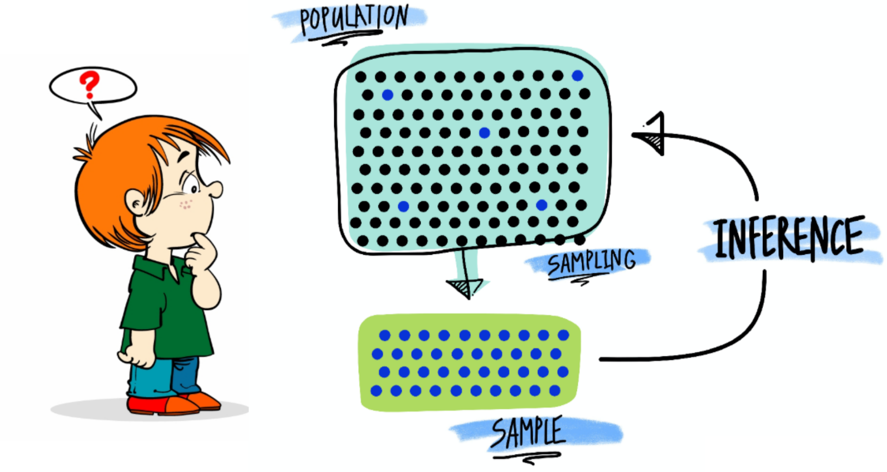
Contents

Part 1:

Hypothesis testing

1. Introductory material

- Objective of statistics is to make **inferences about unknown population parameters** based on information contained in sample data.



These **inferences** are phrased as

▷ estimates of the respective parameters

▷ tests of hypotheses about their values.

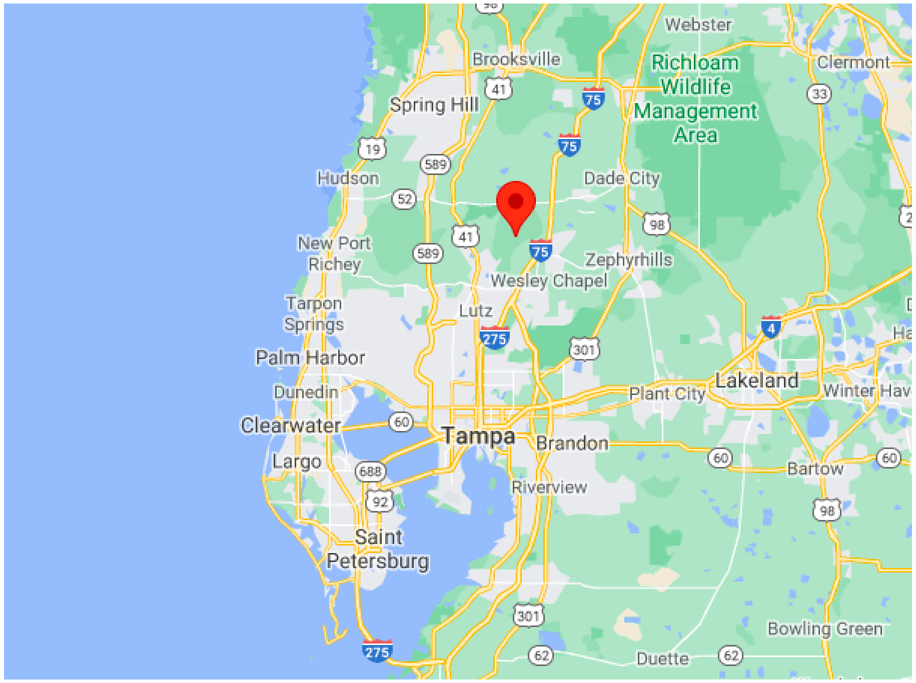
- Hypothesis tests are constructed in **all fields** in which theory can be tested against observation:
  - ▷ A **medical researcher** may hypothesize that a new drug is more effective than another in combating a disease.
  - ▷ A **quality control engineer** may hypothesize that new assembly method produces only 5% defects.
  - ▷ An **educator** may claim that two methods of teaching are equally effective.
  - ▷ A **political candidate** may claim that a plurality of voters favor his election.

  All such hypotheses can be subjected to statistical verification by using observed sample data.

2. Example

Example (from the medical/economic field): Tampa Bay area survey

- The Tampa Bay area is a major populated area surrounding Tampa Bay on the west coast of Florida, US



- A survey was conducted on the Tampa Bay health managers (HMs), with primary goal to investigate the cost effectiveness of these HMs
- Research questions:
  - Who are the Tampa Bay HMs?
  - What is the cost effectiveness of the HMs?

- The (overall) mean of average monthly cost per patient in the US is $\$130$.
- **Research question:** Is the mean cost for family practitioner (fp) different than $\$130$?

The procedure to decide whether there is sufficient evidence to believe the mean cost for family practitioner is different than $\$130$ is called **test of hypothesis**!

To make a conclusion, we will make use of the Tampa Bay area survey, i.e., the observed sample data.

2.1 Null and alternative hypothesis

- In practice, the research question is formulated in terms of a **null hypothesis $\mathbf{H_0}$** and an **alternative hypothesis $\mathbf{H_a}$**:

$$\boxed{H_0 : \mu_{fp} = 130 \qquad H_a : \mu_{fp} \neq 130}$$

- Based on our observed data, we will investigate whether H_0 can be rejected in favour of H_a

- If not, the null hypothesis H_0 is accepted, i.e., fail to reject H_0, and one decides that the mean cost for family practitioner is not different than $\$130$.

- **Question:** But how do we choose/formulate the null (and alternative) hypotheses?

- **Answer:** The choice typically depends on reasons of custom and convenience.
  - It is conventional to choose the simpler of two hypotheses as the null, e.g.,

    H_0 : The distribution is Poisson $\qquad$ H_a : The distribution is not Poisson

  - Consequences of incorrectly rejecting one hypothesis may be graver than those of incorrectly rejecting the other. The former should be chosen as the null, e.g.,

    H_0 : new drug gives same result $\qquad$ H_a : new drug is superior

  - Null hypothesis is often a simple explanation that must be discredited in order to demonstrate the presence of some physical phenomenon or effect. (This must be conclusively disproved in order to convince skeptic that H_a is true.)

- Accepting H_0 (i.e., fail to reject H_0) is a weak conclusion, whereas accepting H_a is a strong conclusion.

Basic strategy in statistics is:

> *Be conservative for H_0 (let H_0 get the advantage of the doubt), and reject the null distribution H_0 if it is no good explanation for the observations* $\mathbf{X}$ *(or better for the observed sample summarized in the sufficient statistic $\overline{X}$), that is if the observation $\overline{X}$ becomes too unlikely under the null distribution.*

2.2 Test statistic

- Intuitively, it is obvious that $H_0 : \mu_{fp} = 130$ will be rejected if the observed sample average $\bar{x}$ is too far away from 130.

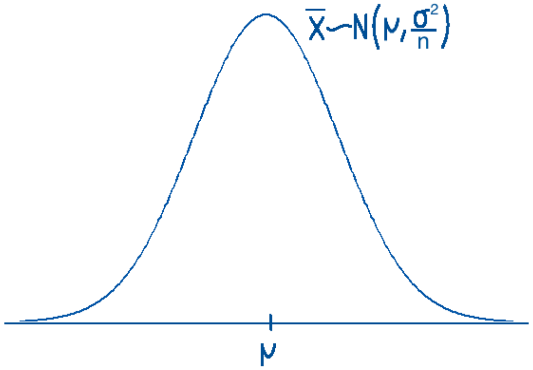
    - Tampa Bay area survey: $n = 50$, $\bar{x} = 121.06$, $s^2 = 83.53$

- **Question:** But what is too far away?

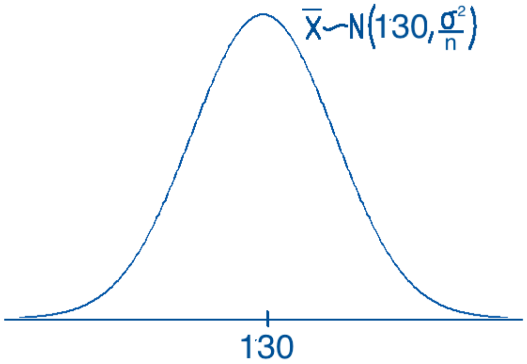
- If this result is very unlikely to happen by pure chance

- Said differently, if this result is not at all what you expect to see if $\mu_{fp} = 130$

- The CLT will help us in deciding, as it describes what values for $\bar{x}$ are to be expected if one would repeatedly draw new samples. If n is sufficiently large, we know that:



- Here, we are interested in knowing what values for $\bar{x}$ could be expected if $\mu_{fp} = 130$



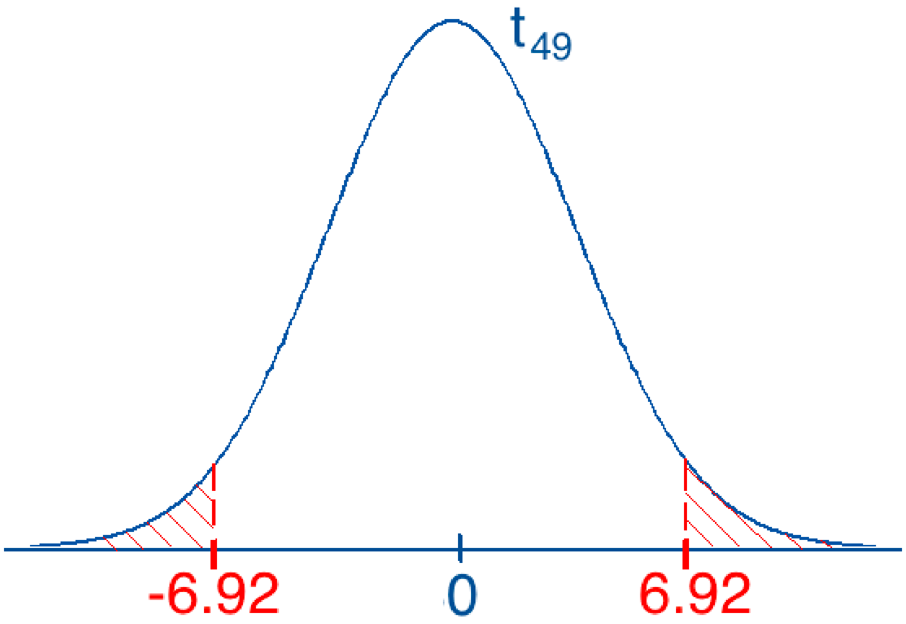
- When σ^2 is unknown, we will base ourself on

$$T = \frac{\bar{X} - 130}{\sqrt{\frac{S^2}{n}}} \overset{H_0}{\sim} t_{df=n-1}$$

  - In hypothesis testing, T is called a **test statistic**
    - In statistics, several test statistics exist, depending on the defined hypothesis (see Section 5)

  - T can be calculated for our given sample, defined as the **sample value (t)**
    - $t = \frac{121.06-130}{1.29} = -6.92$

- To reject H_0, the red area needs to be sufficiently low

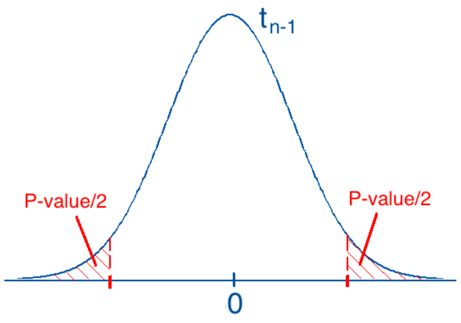


- In statistics, the red area is defined as the **P-value**

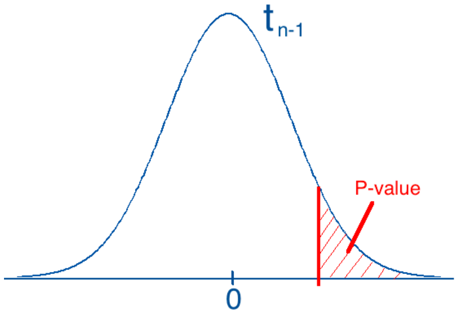
- **P-value** $=$ **Probability** that, if H_0 is true, the **test statistic** is as **extreme as or more extreme than** the **sample value**

- Depending on the hypothesis, different locations are given for this value

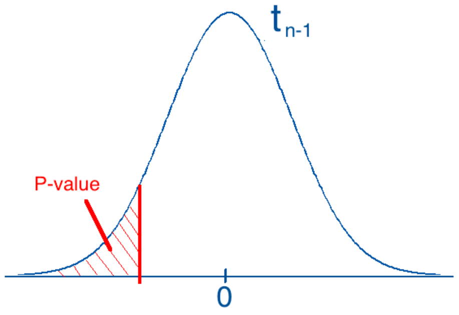
  - $\boxed{H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0}$



  - $\boxed{H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0}$



  - $\boxed{H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0}$



- **Question:** But what is sufficiently low?

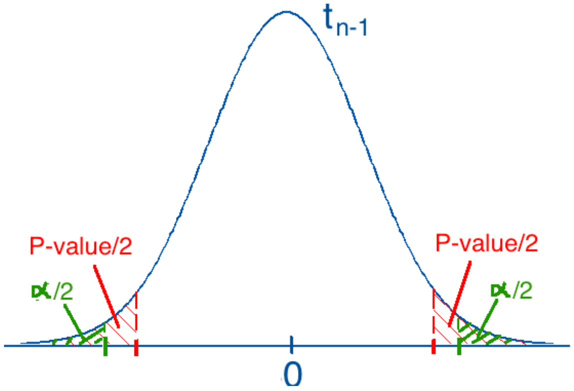
- One therefore specifies the so-called level of significance α
  - $\alpha =$ **Probability that**, if H_0 is true, the **test statistic** is not able to detect this

  - This means that the confidence level $(1-\alpha)$ defines the probability that, if H_0 is true, the test is able to detect this

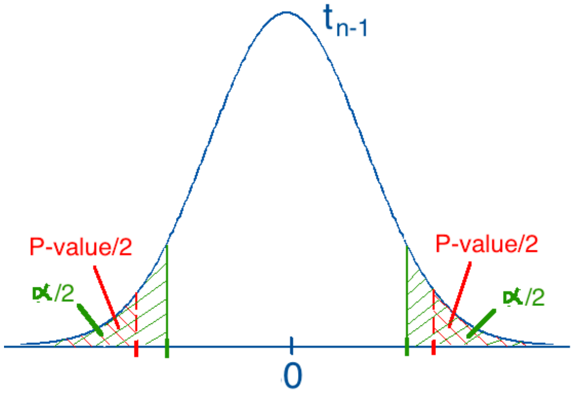
  - In research, one often use 1%, 5% of 10% for α

2.3 Decision making

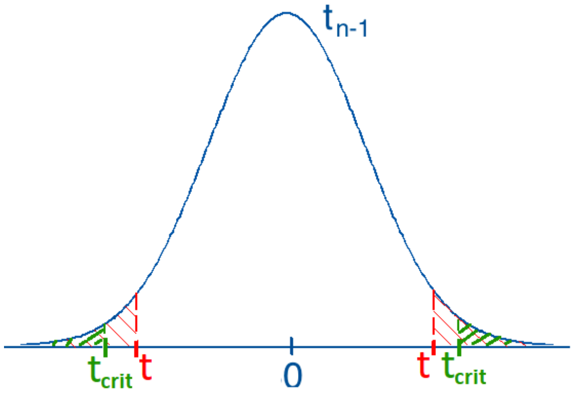
- **Decision (based on the P-value):**
  - Accept H_0 if $P \geq \alpha$



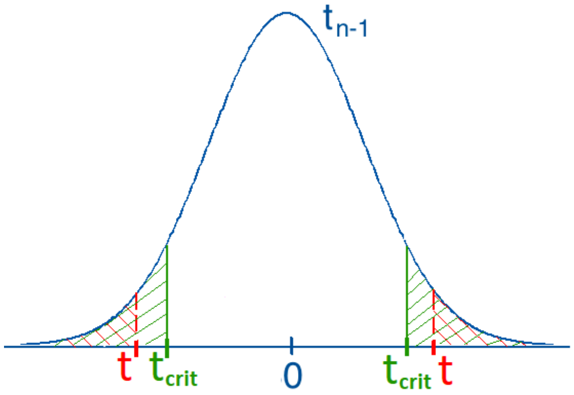
  - Reject H_0 if $P < \alpha$



- **Decision (based on the sample value):**
  - Accept H_0 if $t \leq t_{crit}$



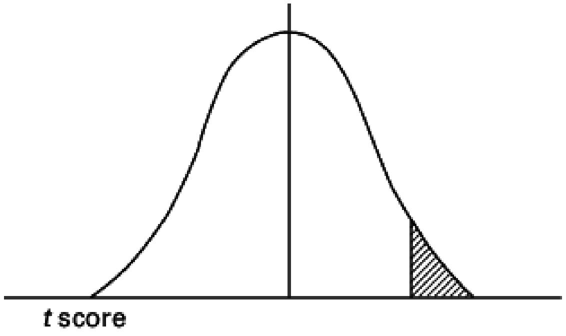
  - Reject H_0 if $t > t_{crit}$

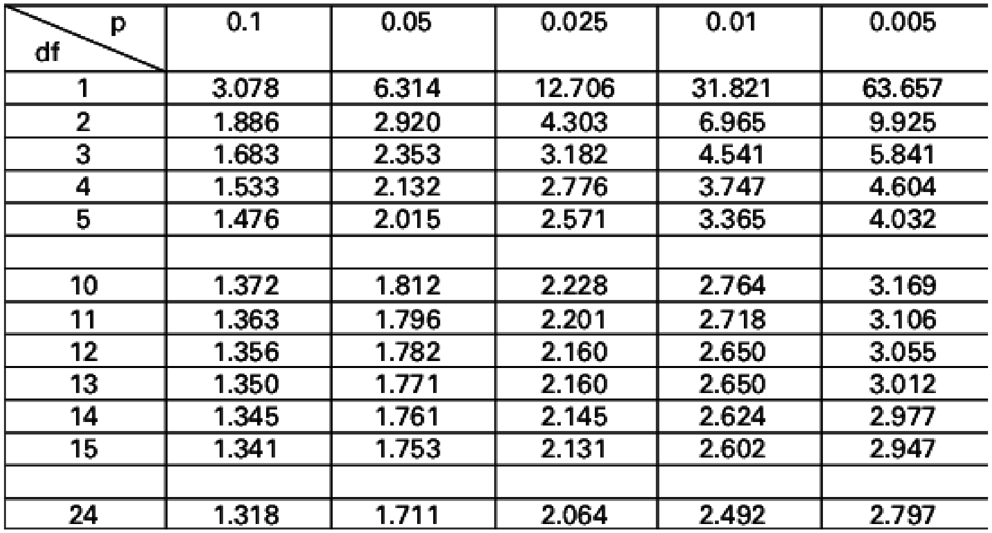


  - **Remark:** The critical value (t_{crit}), of course, depends on the chosen α.

- **Question:** But where can we find this critical value t_{crit}?

- **Answer:** Tables





p df	0.1	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	1.683	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
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	2.160	2.650	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
24	1.318	1.711	2.064	2.492	2.797

2.4 Exercises

Exercise 1:

Actual production of bulbs has lifetimes with mean $830h$ and variance $(40h)^2$ (assume that the lifetimes follow a normal distribution). A new production will be started if it gives bulbs with a higher lifetime. To investigate whether the new production will be started or not, a random sample of size 10 is taken, with measures $n = 10, \bar{X} = 845, s = 40$. Should the new production be started?

Solution exercise 1:

1. Variable X: lifetime of random bulb in new production.
2. Parameter of interest: μ.
3. Consider significance level $\alpha = 5\%$.

4. Hypothesis testing:

$$\boxed{H_0 : \mu \leq 830 \qquad H_a : \mu > 830}$$

5. Intuitive test: Reject H_0 if $\bar{X}$ is too large.

6. Test statistic: $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$.

7. Sample value: $z = \frac{845-830}{40/\sqrt{10}} = 1.186$.

8. P-value: $p = P(Z \geq z \mid H_0\ true) = 0.118 > \alpha$.

9. Conclusion: Accept H_0 (reject H_a), i.e., keep the old production.

The argument can be phrased as follows:

- Data are consistent with null hypothesis, for they are quite likely under the null distribution: Under mean lifetime 830 an observed mean as extreme as the observed 845 will occur in somewhat 11.8% (about 1 in 8) of the samples.

- Data do not provide evidence against the null hypothesis.

- Difference can be explained by random fluctuation: Due to variation in lifetimes ($\sigma = 40$) such a difference is quite likely.

3. Types of errors

		Reality	
		H_0 correct	H_0 not correct
Test result	H_0 correct	OK	**Type II error**
	H_0 not correct	**Type I error**	OK

- **Type I error:**
  - Occurs if H_0 is correct but the test leads to a significant result
  - **Question:** What is the chance that this occurs?
  - Suppose the test is performed at $\alpha = 5\%$
  - If H_0 is correct, then one will observe a significant result in 5% of the cases

- Thus, P(Type I error) $= \alpha$

- **Type II error:**
  - Occurs if H_0 is incorrect but the test has not detected this
  - **Question:** What is the chance that this occurs?
  - In contrast to the type I error, the probability of making a type II error is not easily controlled, and depends on various aspects of the sample(s) and population(s), and is denoted by β
  - The **power** of a statistical test is $\mathbf{1} - \beta$, the probability of correctly rejecting H_0

- **Be aware:** β can not be explicitly computed because 'H_0 is not true' gives us no information about unknown parameter of interest, unless a specific alternative value is given, for example, $H_a : \mu = \mu_1 \neq \mu_0$.

  Example: Tampa Bay area survey

  $$\begin{aligned} P(\text{type II error}|\mu = \mu_1) &= P(\text{accept } H_0|\mu = \mu_1 \text{ true}) \\ &= P\left(-t_{n-1,\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_{n-1,\frac{\alpha}{2}} \mid \mu = \mu_1\right) \end{aligned}$$

  for $\mu = \mu_1$ it holds that $T = \frac{\bar{X}-\mu_1}{S/\sqrt{n}} \sim t_{n-1}$ hence

  $$\begin{aligned} &= P\left(-t_{n-1,\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \leq \frac{\bar{X} - \mu_1}{S/\sqrt{n}} \leq t_{n-1,\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{S/\sqrt{n}}|\mu = \mu_1\right) \\ &= P\left(-t_{n-1,\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \leq T \leq t_{n-1,\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{S/\sqrt{n}}\right) \end{aligned}$$

- This gives rise to **power analysis**!

3.1 Power

- In general, a specific testing procedure is acceptable, only if:
  - The chance of making a type I error is sufficiently small;
  - The power to detect deviations from H_0 is sufficiently large.
- The first condition can easily be met by specifying α sufficiently small.
- The second condition is more difficult to meet, as the power depends on various aspects of the sample(s) and population(s).
- In what follows, this will be illustrated in the context of the comparison of 2 independent groups.

SETTING:

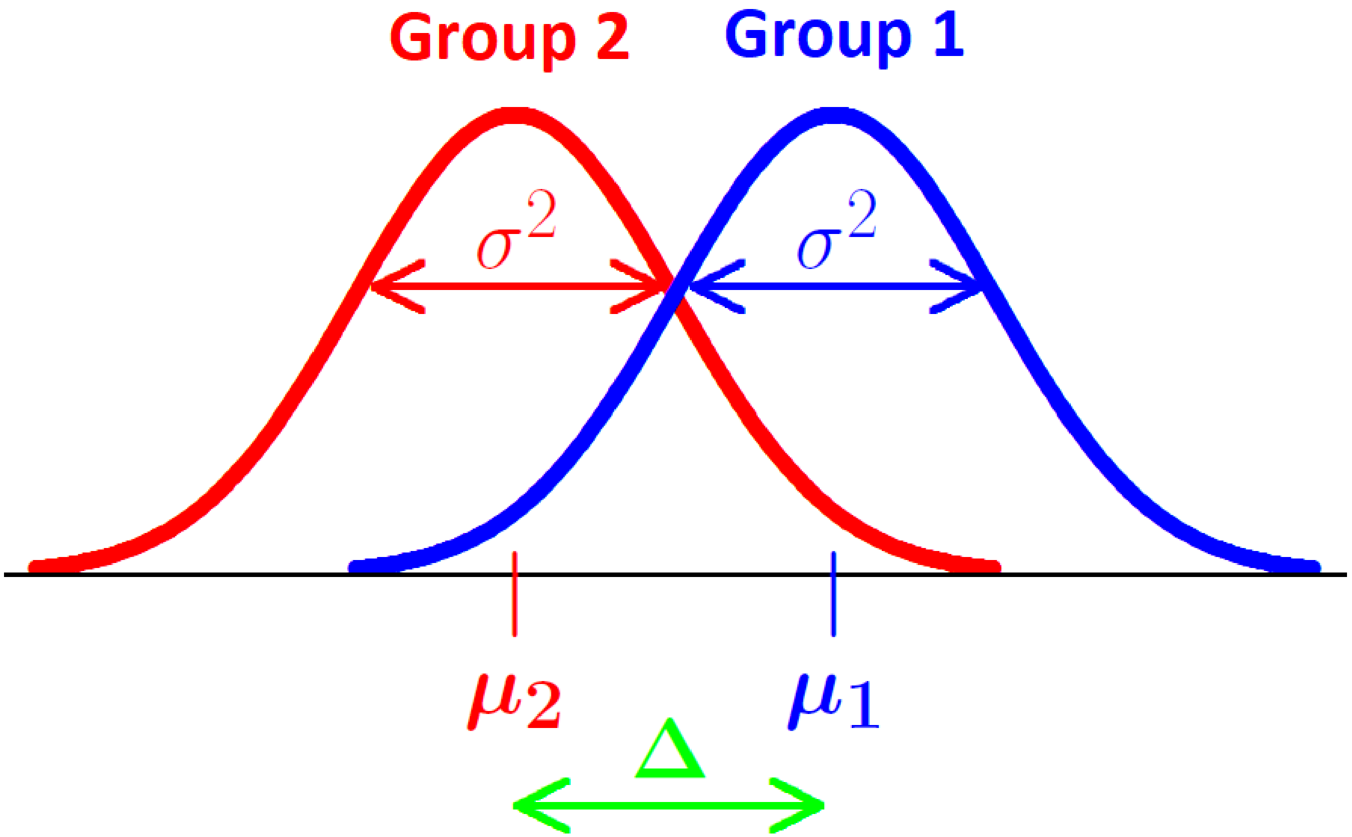
- The null and alternative hypotheses are given by

$$\boxed{H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2.}$$

- In the case H_a is true, i.e., $\mu_1 \neq \mu_2$, we denote the true difference between populations by $\Delta = \mu_1 - \mu_2$.

- Assume the data to be normally distributed in both populations, with equal variability σ^2.

- Graphically:



3.1.1 Power as a function of the significance level α

The smaller α, the smaller the power

- **Intuitively:** Type I errors are less likely if the null hypothesis is rejected less often. However, in cases where H_0 is truly wrong, it will still be rejected less often.

3.1.2 Power as a function of the sample size(s)

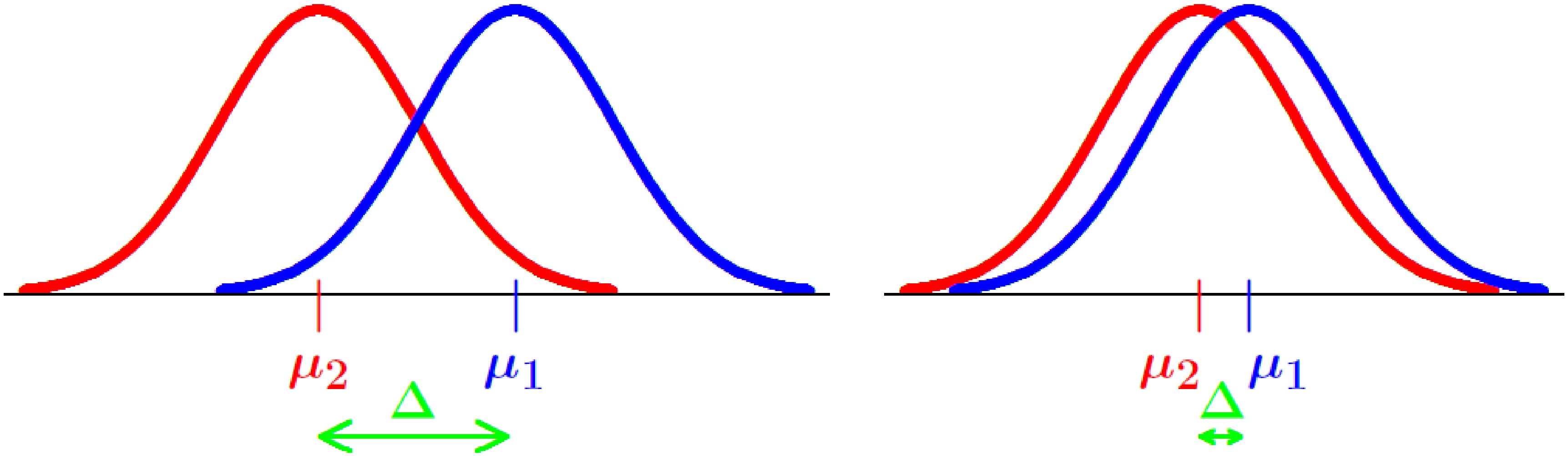
The more observations, the larger the power

- **Intuitively:** More observations yields more information about the population(s), therefore implying more precision in the conclusions.

3.1.3 Power as a function of the true difference $\triangle$

The smaller $\triangle$, the smaller the power

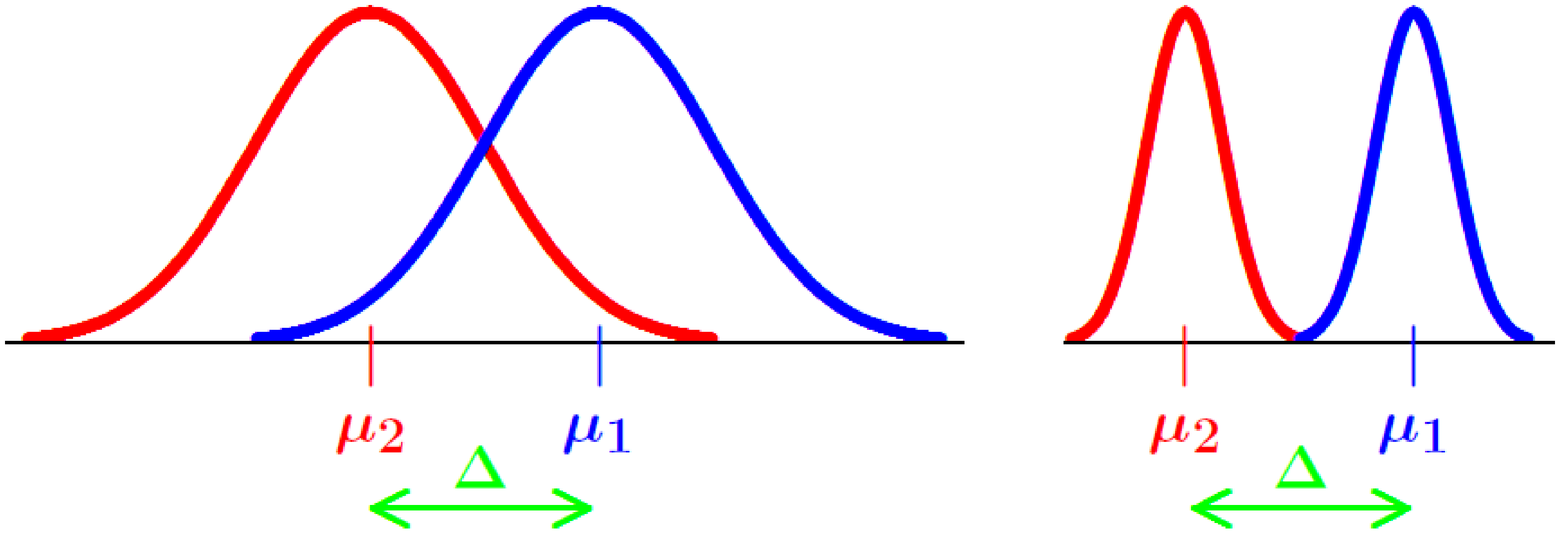
- **Intuitively:** Large deviations from the null hypothesis are easier to detect



3.1.4 Power as a function of the variability σ^2

The smaller σ^2, the larger the power

- **Intuitively:** Large deviations from the null hypothesis are easier to detect



- The power depends on various aspects:
  - The significance level α;
  - The true difference Δ between the populations;
  - Within-group variance σ^2;
  - The sample size(s).
- **Visualization:** $http : //rpsychologist.com/d3/nhst/$
- In what follows, we will illustrate power calculation with some examples.

Exercise 1:

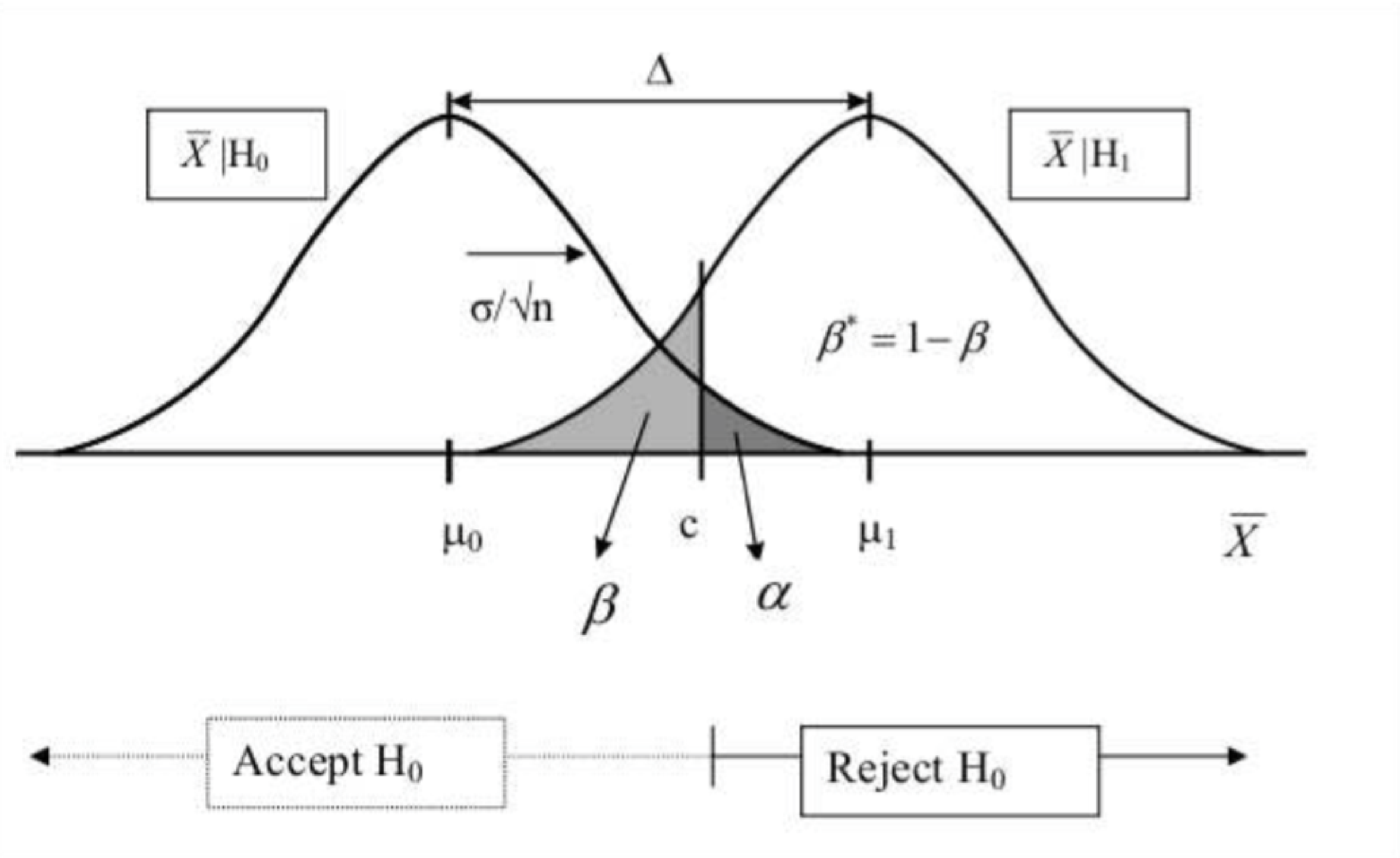
Reconsider the bulb lifetime problem, i.e.,

- Variable X: Lifetime of random bulb in new production;
- $X \sim N(\mu, \sigma^2)$, with $\sigma = 40$;
- Observed data: $n = 10, \bar{X} = 845$;
- Hypothesis testing:

$$\boxed{H_0 : \mu \leq 830 (= \mu_0) \qquad H_a : \mu > 830}$$

A researcher considers following test: Reject H_0 if $\bar{X} > 842 (= c)$.

Find the significance level of this test and its power for alternative $\mu = 850 (= \mu_1)$.


$\bar{X}\,|H_0$
Δ
$\bar{X}\,|H_1$
$\sigma/\sqrt{n}$
$\beta^* = 1 - \beta$
μ_0
c
μ_1
$\bar{X}$
β
α
Accept H_0
Reject H_0

Solution exercise 1:

From CLT:

$$\overline{X} \sim N\left(\mu, \frac{40^2}{10}\right) \sim N(\mu, 12.65^2).$$

Significance level of the test:

$$\alpha = P(\text{reject } H_0 | H_0 \text{ true}) = P(\overline{X} > 842 | \mu = 830) = 0.171.$$

Interpretation:

- If the new production is not better, the test will nevertheless conclude in 17,1% of the samples that it is better.

Power of the test for the alternative $\mu = 850$:

$$\beta^*(850) = P(\text{reject } H_0|H_1 \text{ with } \mu = 850) = P(\overline{X} > 842|\mu = 850) = 0.736$$

Interpretation:

- If the new production is better ($\mu = 850$), then the test will detect it in 73.6% of the samples.

Exercise 2:

Medical researchers are interested in investigating the effect of medical drug on increasing iron level in blood of teenage girls. They have gathered the following information:

- Variable X: Increases in iron level (in $\mu g/dL$);
- $X \sim N(\mu, \sigma^2)$, with $\sigma = 40$;
- Observed data: $n = 20, \bar{X} = 18$;
- Hypothesis testing:

$$\boxed{H_0 : \mu \leq 0 \qquad H_a : \mu > 0}$$

- Significance level: $\alpha = 5\%$

Calculate the power at alternative $\mu = 20$.

Solution exercise 2:

From CLT:

$$\overline{X} \sim N\left(\mu, \frac{40^2}{20}\right) \sim N(\mu, 8.94^2).$$

Critical value (= c) at $\alpha = 5\%$:

$$\alpha = P(\text{reject } H_0|H_0 \text{ true}) = P(\overline{X} > c \mid \mu = 0) = 0.05.$$

$$\Leftrightarrow$$

$$\frac{c}{8.94^2} = 1.645 \Rightarrow c = 14.7$$

Power of the test for the alternative $\mu = 20$:

$$\beta^*(20) = P(\text{reject } H_0|H_1 \text{ with } \mu = 20) = P(\overline{X} > 14.7|\mu = 20) = 0.723$$

Interpretation:

- If the drug has effect $\mu = 20$, then about 72% of the samples will show effect in the test.

3.2 Sample size calculation

- **Remark:** In power analysis, the sample size is the only aspect under control of the investigator.

- In practice, e.g., clinical trials, one can calculate the sample size needed to reach a sufficiently high power, leading to so-called **sample size calculations**, i.e., a specific part of power analysis.

- To conduct this sample size,
  - The level of significance needs to be chosen apriori;
  - The within-group variance σ^2 needs to be pre-specified based on earlier, similar experiments, relevant literature, or a pilot study
  - In practice, Δ is not known. Instead, the smallest Δ which would still be practically relevant to detect, needs to be specified.

SETTING:

- Assume the following null and alternative hypotheses:

$$\boxed{H_0 : \mu \leq \mu_0 \qquad H_a : \mu > \mu_0.}$$

- Let $X \sim N(\mu, \sigma^2)$, with σ known;
- Test statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0, 1)$.
- Significance level α

PROCEDURE:

- Find the rejection region for H_0: Reject H_0 if $\bar{X} > \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$, where z_α is $\alpha-$upper quantile of standard normal distribution.

- Construct the power of the test at μ_1:

$$\beta^*(\mu_1) = P\left(Z > z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right), \text{ where } Z \sim N(0,1)$$

.

- Find sample size n such that the power at μ_1 is at least a level β^*:

$$\beta^*(\mu_1) \geq \beta^* \qquad \Leftrightarrow \qquad z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq z_{\beta^*}$$

$$\boxed{n \geq \left[\tfrac{\sigma}{\mu_1 - \mu_0} \cdot (z_\alpha - z_{\beta^*})\right]^2}$$

- **Remark:** A similar approach can be obtained for other hypothesis testing!

Exercise 3:

Concentration of chemical in preparation is assumed to be 20 units. However, the medium may have been interchanged with previous preparation of concentration 24. Experimenter plans to check for concentration 20. He will

- measure concentration a number of times, n.
- observe $\overline{X}$
- perform a classical statistical z-test:
  Reject concentration $\mu_0 = 20$ if

$$\overline{X} > \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}},$$

where z_α is $\alpha-$upper quantile of standard normal distribution.

Assume measuring method has standard deviation $\sigma = 3$.

How many measurements should be performed in order that the test has a type I error probability of at most 5% and a power of at least 90% to detect the 'wild' concentration 24?

Solution exercise 3:

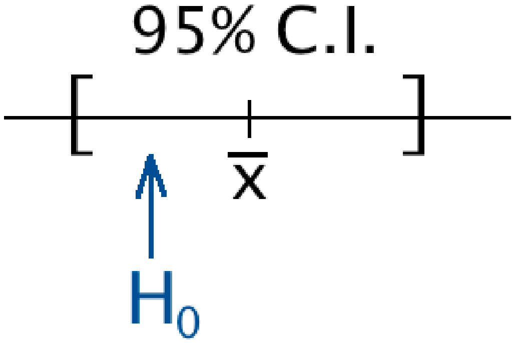
$$n \geq \left[\frac{3}{24 - 20}(1.6449 + 1.2816)\right]^2 = 4.8 \quad \text{or} \quad n = 5.$$

4. Hypothesis testing versus confidence intervals

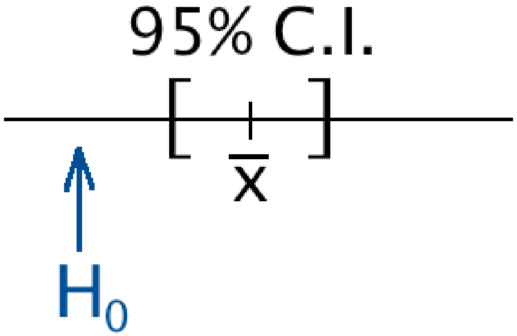
- For the **Tampa Bay area doctors survey**, we have drawn conclusions about the population average cost per patient per month
  - $H_0 : \mu_{fp} = 130$ & $H_a : \mu_{fp} \neq 130 \rightarrow p < 0.00001$
  - 95% C.I.: $[44.51; 82.98]$
- We know from the C.I. that the average average cost is likely to be between 44.51 and 82.98, excluding 130
- The significance test has rejected the value 130 as possible value for μ_{fp}
- So, **both procedures agree!**

- **Question:** But does this always agree?

- **Answer:** Yes, provided the levels of significance and confidence are **complementary to each other**, e.g.,
  - Accept H_0 ($p \geq \alpha = 0.05$:)



  - Reject H_0 ($p < \alpha = 0.05$:)



THEOREM:

Consider a parameter θ and a statistic $T = T(X)$ for θ. Then a $(1 - \alpha)$ confidence interval for θ is the set of acceptable null hypotheses at level α. (Simple null hypothesis against the total alternative.)

PROOF:

Case of normal mean & two-sides hypothesis testing (general case is similar):
$X \sim N(\mu, \sigma^2)$, with σ known, and hypothesis $H_0 : \mu = \mu_0$ and $H_a : \mu \neq \mu_0$.

- Test statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

- Conclusion: Reject H_0 if $\bar{X} > \mu_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ or $\bar{X} < \mu_0 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$$\text{Accept } H_0 \Leftrightarrow \mu_0 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \mu_0 \text{ belongs to the } (1-\alpha)100\% \text{ C.I. for } \mu$$

Part 2:

Different tests

1. Classical tests

- In practice, hypothesis testing is done for **means**, **proportions** and **variances**, leading to a different set of test statistics.

- In what follows, we will give an overview of some classical tests for these situations.

- **Remark:** The construction of hypothesis test is similar to the construction of corresponding C.I., but here we derive the distribution of the test statistic under the null hypothesis.

Hypothesis	Assumptions	Test statistic and distribution under H_0
TEST FOR ONE MEAN		μ
$H_0 : \mu = \mu_0$	Normal population known variance σ^2	**z-test** $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$
	Normal population unknown variance	**t-test** $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$
	General population large sample (CLT)	**z-test** (approx.)[1] $Z = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1)$

[1] If $n \to \infty$: $t_{n-1} \approx N(0,1)$

Hypothesis	Assumptions	Test statistic and distribution under H_0
TEST FOR TWO MEANS GIVEN INDEPENDENT SAMPLES μ_1 and μ_2		
$H_0 : \mu_1 = \mu_2$	Normal populations known variances	**z-test** $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0, 1)$
	Normal populations unknown variances $\sigma_1^2 = \sigma_2^2$	**t-test** $T = \dfrac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
	Normal populations unknown variances $\sigma_1^2 \neq \sigma_2^2$	**t-test** $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_\nu$ $\nu \approx \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$

Hypothesis	Assumptions	Test statistic and distribution under H_0

TEST FOR TWO MEANS GIVEN PAIRED DATA μ_1 and μ_2

$H_0 : \mu_1 = \mu_2$	Test for equal means μ_1, μ_2	**t-test**
	for paired data (X_i, Y_i).	$T = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1}$
equivalently	Difference data $D_i = X_i - Y_i$.	$\overline{D} = \frac{\sum D_i}{n} = \overline{X} - \overline{Y}$
	Test for mean 0	$S_D^2 = \frac{\sum (D_i - \overline{D})^2}{n-1}$
$H_0 : \mu_D = 0$	D normally distributed	

Hypothesis	Assumptions	Test statistic and distribution under H_0

TEST FOR ONE PROPORTION	p

$H_0 : p = p_0$	Observed proportion in binomial sample: $\hat{P} = \sum X_i / n$	**z-test** (approx.) $Z = \dfrac{\hat{P} - p_0}{SE(\hat{P})} \approx \mathrm{N}(0,1)$ $SE(\hat{P}) = \sqrt{\frac{p_0(1-p_0)}{n}}$

TEST FOR TWO PROPORTIONS	p_1 and p_2

$H_0 : p_1 = p_2$	Comparison of proportions p_1, p_2 in 2 independent samples. Observed proportions $\hat{P}_1 = \sum X_{1i}/n_1,\ \hat{P}_2 = \sum X_{2i}/n_2$	**z-test** (approx.)[2] $Z = \dfrac{\hat{P}_1 - \hat{P}_2}{\widehat{SE}(\hat{P}_1 - \hat{P}_2)} \approx \mathrm{N}(0,1)$ $\widehat{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1} + \frac{1}{n_2})}$ $\hat{P} = \dfrac{\sum X_{1i} + \sum X_{2i}}{n_1 + n_2} = \dfrac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2}$

[2] See Section 5.1

Hypothesis	Assumptions	Test statistic and distribution under H_0

TEST FOR A VARIANCE		σ^2
$H_0 \ : \ \sigma^2 = \sigma_0^2$	Normal population	χ^2 **test** $\chi^2 = (n-1)\,\dfrac{S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

TEST FOR TWO VARIANCES		σ_1^2 and σ_2^2
$H_0 \ : \ \sigma_1^2 = \sigma_2^2$	Independent normal populations	**F-test** $F = \dfrac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$

1.1 Test statistic for two proportions

Let $H_0 : p_1 = p_2$ and assume $p = p_1 = p_2$.

Denote Q_1 and Q_2 respectively the number of successes in group 1 and 2 (with sample size n_1 and n_2).

$$Q_1 \sim B(n_1, p) \ \approx \ N(n_1 p, n_1 p(1-p)) \Rightarrow \hat{p}_1 = \frac{Q_1}{n_1} \sim N\left(p, \frac{p(1-p)}{n_1}\right)$$

$$Q_2 \sim B(n_2, p) \ \approx \ N(n_2 p, n_2 p(1-p)) \Rightarrow \hat{p}_2 = \frac{Q_2}{n_2} \sim N\left(p, \frac{p(1-p)}{n_2}\right)$$

Thus,

$$\hat{p}_1 - \hat{p}_2 \ \approx \ N\left(0, \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}\right)$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \approx N(0,1)$$

Since p is unknown, this needs to be estimated: $\hat{p} = \frac{Q_1+Q_2}{n_1+n_2} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$

1.2 Exercises

Exercise 1:

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 and a sample standard deviation of 5.

Can we conclude at the 0.05 significance level that the two unknown population variances differ?

Solution exercise 1:

Let σ_1^2 and σ^2 respectively be the population variances for the abrasive wear of material 1 and material 2.

- $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$.
- $f = \frac{s_1^2}{s_2^2}$ with $v_1 = 11$ and $v_2 = 9$.
- Critical region $(\alpha = 005)$: $F_{11,9,0.975} = 3.91$ and $F_{11,9,0.025} = \frac{1}{F_{9,11,0.975}} = 0.28$. $\Rightarrow H_0$ rejected when $f < 0.28$ or $f > 3.91$.
- Sample value: $f_{obs} = \frac{16}{25} = 0.64$.

Conclusion: Do not reject H_0. There is insufficient evidence that variances differ.

Exercise 2:

We want to measure the yield of a chemical process according to the used catalytic converter (A or B). A sample of 5 products using A resulted in a sample mean of 20.84 and sample standard deviance of 7.246, whereas a sample of 6 products using B resulted in a sample mean of 22.53 and sample standard deviance of 5.432.

Are the yields different on a significance level $\alpha = 0.05$?

Solution exercise 2:

- $H_0 : \mu_A = \mu_B$ versus $H_a : \mu_A \neq \mu_B$.
- Test the ratio of the variances:

$$f = \frac{52.50}{29.10} = 1.78 \in \left[\frac{1}{F_{5,4,0.975}}, F_{4,5,0.975}\right] = [1/9.63, 7.39]$$

$\Rightarrow$ We accept $\sigma_1 = \sigma_2$

- Pooled variance:

$$s^2 = \frac{4 \cdot 52.20 + 5 \cdot 29.51}{4+5} = 39.73 \quad \Rightarrow \quad s = 6.3$$

- Test:

$$|t| = \frac{|20.84 - 22.53|}{6.3 \cdot \sqrt{\frac{1}{5} + \frac{1}{6}}} = 0.443 < t_{5+6-2;0.975} = 2.262$$

Conclusion: Do not reject H_0. There is insufficient evidence that different catalytic converters give different yields.

Exercise 3:

We test the roughness of metal surfaces after polishing up using two different milling machines. Each metal test piece is splitted in two and both splices are handled to the two different machines, giving the following table (column 2 and 3 are measures of roughness for both machines).

test piece	machine 1	machine 2
1	1.77	1.56
2	1.32	1.30
3	1.83	1.85
4	1.02	0.94
5	1.92	1.95
6	1.68	1.55
7	2.48	2.60
8	1.55	1.32
9	0.99	0.85
10	1.96	2.00

Is there a difference between both methods (significance level $\alpha = 0.05$)?

Solution exercise 3:

Paired data!

- $H_0 : \mu_D = 0$ versus $H_a : \mu_D \neq 0$.
- Results for the differences D_i:

$$\begin{aligned} \overline{d} &= 0.06 \\ s_D &= 0.1162 \end{aligned}$$

- Test:

$$|t| = \frac{|0.06 - 0|}{0.1162/\sqrt{10}} = 1.663 < t_{9;0.975} = 2.262$$

Conclusion: Do not reject H_0. There is insufficient evidence that different methods yield different roughness.

Exercise 4:

A vote is to be taken among the residents of a town and the surrounding county to determine whether a proposed chemical plant should be constructed. The construction site is within the town limits, and for this reason many voters in the county feel that the proposal will pass because of the large proportion of town voters who favor the construction. To determine if there is a significant difference in the proportions of town voters and county voters favoring the proposal, a poll is taken.

If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use an $\alpha = 0.05$ level of significance.

Solution exercise 4:

Let p_1 and p_2 be respectively the true proportions of voters in the town and county favoring the proposal.

- $H_0 : p_1 \leq p_2$ versus $H_a : p_1 > p_2$.
- Critical region: $z > 1.645$ $(\alpha = 0.05)$.
- Computations

$$\hat{p}_1 = \frac{q_1}{n_1} = 0.60; \quad \hat{p}_2 = \frac{q_2}{n_2} = 0.48; \quad \hat{p} = \frac{q_1 + q_2}{n_1 + n_2} = 0.51.$$

- Test statistic: $z = \frac{0.60 - 0.48}{\sqrt{(0.51)(0.49)(1/200+1/500)}} = 2.9.$
- $P(Z > 2.9) = 0.0019.$

Conclusion: Reject H_0 and agree that proportion of town voters favoring proposal is higher than proportion of county voters.

2. MLE based tests

- Consider a one-dimensional parameter θ, and it's ML estimator $\hat{\theta}_{ML}$.

- Suppose we are interested in the following hypotheses:

$$\boxed{H_0 : \theta = \theta_0 \qquad H_a : \theta \neq \theta_0}$$

▷ **Reminder (Chapter 7):** Under regularity conditions, we have for a one-dimensional parameter θ:

$$\sqrt{nI(\theta)} \cdot \left(\hat{\theta}_{ML} - \theta\right) \approx N(0,1), \text{ for } n \text{ sufficiently large.}$$

- Thus, under H_0, we have:

$$\sqrt{nI(\theta_0)} \cdot \left(\hat{\theta}_{ML} - \theta_0 \right) \approx N(0,1),$$

leading to

$$\text{P-value} = 2 \cdot P(Z > |z_{obs}|),$$

with $Z = \sqrt{nI(\theta_0)} \cdot \left(\hat{\theta}_{ML} - \theta_0 \right).$

2.1 Example: Poisson

Consider $X \sim Pois(\lambda)$, $P(X = x) = \frac{\lambda^x e^{-x}}{x!}$, $x = 0, 1, \ldots$.

- **ML estimator:** $\hat{\lambda} = \bar{X}$
- **Fisher information:** $I(\lambda) = \frac{1}{\lambda}$
- Suppose we are interested in the following **hypotheses**:

$$\boxed{H_0 : \lambda \leq \lambda_0 \qquad H_a : \lambda > \lambda_0}$$

- Under H_0, we have

$$\sqrt{n\lambda_0} \cdot \left(\frac{\bar{X}}{\lambda_0} - 1\right) \approx N(0, 1),$$

- **Conclusion:** Reject H_0 when $\bar{X} > \lambda_0 + z_\alpha \cdot \sqrt{\frac{\lambda_0}{n}}$

3. Likelihood ratio tests

- Let the total parameter set be denoted by Θ and the subset corresponding to H_0 by Θ_0

- The **likelihood ratio (LR)** test is then defined by

$$\Lambda = \frac{\text{max}_{\theta\in\Theta_0} L(\theta)}{\text{max}_{\theta\in\Theta} L(\theta)}.$$

- **Conclusion:** Reject H_0 for too small values of Λ, or equivalently, for too large values of $-2\text{log}\Lambda$.

- $\boxed{\text{THEOREM:}}$

Under regularity assumptions (not shown here), and $n \rightarrow \infty$:

$$-2\log\Lambda \xrightarrow{D} \chi^2_{\mathrm{dim}(\Theta)-\mathrm{dim}(\Theta_0)}.$$

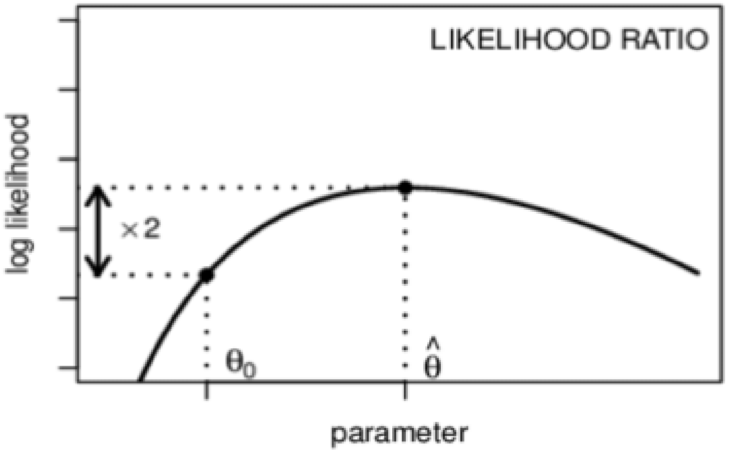
3.1 Example: Exponential

- Let $X_1, \ldots, X_n$ be i.i.d. observations from the exponential model with density

$$\theta e^{-\theta x}, \qquad x > 0, \theta > 0.$$

- **Task:** Derive the LR test of approximate level α (for large sample size) for the hypothesis problem

$$H_0 : \theta = \theta_0 \qquad vs. \qquad H_a : \theta \neq \theta_0.$$



- Derivation of the MLEr $\hat{\theta}$:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta e^{-\theta X_i} \\ \text{log} L(\theta) &= n \text{log} \theta - \theta \sum_{i=1}^{n} X_i \\ \frac{\partial}{\partial \theta} \text{log} L(\theta) &= \frac{n}{\theta} - \sum_{i=1}^{n} X_i. \end{aligned}$$

Thus, the MLEr is given by $\hat{\theta} = 1/\bar{X}$.

- The LRT for $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ is given by

$$\begin{aligned}\Lambda &= \frac{\theta_0^n e^{-\theta_0 \sum_{i=1}^n X_i}}{(\bar{X})^{-n} e^{-\sum_{i=1}^n X_i/\bar{X}}} \\ &= (\theta_0 \bar{X})^n e^{-n\theta_0 \bar{X}+n} \\ &= \left(\theta_0 \bar{X} e^{-\theta_0 \bar{X}+1}\right)^n.\end{aligned}$$

and

$$-2\text{log}\Lambda = -2n\left[\text{log}(\bar{X}\theta_0) + 1 - \bar{X}\theta_0\right] \xrightarrow{D} \chi_1^2.$$

as $n \to \infty$ if $H_0 : \theta = \theta_0$ holds true.

- PROPERTY:

The LRT is asymptotically equivalent to the MLE based test.

▷ **Illustration based on the example:**

∗ **Reminder:** The LRT for $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ is given by

$$-2\log\Lambda = -2n\left[\log(\bar{X}\theta_0) + 1 - \bar{X}\theta_0\right]$$

∗ Under H_0, we know (thanks to the LLN) that $\bar{X}\theta_0 \xrightarrow{P} 1$.

∗ Thus,

$$\begin{aligned} \log(\bar{X}\theta_0) &= \log\left[1 - (1 - \bar{X}\theta_0)\right] \\ &\approx -(1 - \bar{X}\theta_0) - \frac{1}{2}(1 - \bar{X}\theta_0)^2 - \ldots \end{aligned}$$

and

$$-2\text{log}\Lambda \approx -2n\left[-\frac{1}{2}(1-\bar{X}\theta_0)^2 - \ldots\right] = \left[\sqrt{n}(1-\bar{X}\theta_0) + \ldots\right]^2.$$

$*$ From likelihood theory, we have, under H_0, that, as $n \to \infty$:

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0,1).$$

$*$ Since $\frac{\partial}{\partial\theta}\text{log}L(\theta) = -n/\theta^2$, we have: $I(\theta) = \theta^-2$

$*$ Hence, under H_0, as $n \to \infty$:

$$\sqrt{n\theta_0^-2}\left(\frac{1}{\bar{X}} - \theta_0\right) = \sqrt{n}(1-\theta_0\bar{X})\frac{1}{\theta_0\bar{X}} \xrightarrow{D} N(0,1).$$

and

$$\sqrt{n}(1 - \theta_0 \bar{X}) \xrightarrow{D} N(0,1).$$

as $\theta_0 \bar{X} \longrightarrow 1$ as $n \to \infty$ under H_0.

- Hence, the LRT is asymptotically ($n \to \infty$) equivalent to the LME based test rejecting H_0 when the P-value

$$2P\left(Z > |\sqrt{n}(1 - \theta_0 \bar{x}_{obs})|\right),$$

with $Z \sim N(0,1)$, is smaller than the significance level α.

4. Tests for categorical data

- Many experiments result in measurements that are **qualitative** or **categorical** rather than quantitative, e.g., recording whether a person has hypertension (Yes/No).

- Data associated with such measurements can be summarized by providing **the count of the number of measurements** that fall into each of the distinct categories associated with the variable, e.g., the number of people who have hypertension and not.

- In the **analysis of count data**, different questions can be asked:
  - ▷ [GOODNESS OF FIT] Does the observed frequency distribution differs from a theoretical distribution (like Poisson, for example)?
  - ▷ [INDEPENDENCY] Are observations consisting of measures on two variables, expressed in a contingency table, independent of each other?
  - ▷ [HOMOGENEITY] Is the distribution of counts for two or more groups using the same categorical variable the same or not?
- **Question:** Is there a test that can answer these questions?
- **Answer:** Yes, the **Pearson's** χ^2 **test**!

- In 1900, Karl Pearson proposed the χ^2 test statistic (denoted here by Q), which is a function of the squares of the deviations of the observed counts from their expected values, weighted by the reciprocals of their expected values, i.e.,

$$Q = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  - ▷ O_{ij}: the observed counts (obtained from a frequency table or contingency table).
  - ▷ E_{ij}: the expected counts under H_0 (need to be estimated under H_0).
  - ▷ I: number of categories for variable 1
  - ▷ J: number of categories for variable 2 (if present)

4.1 Goodness of fit test for counts

- Suppose we have a frequency distribution available from the observed sample.
  Example:

n	O_i
0-2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9
17+	5

- **Research question:** Is the Poisson (or any other count) distribution plausible here?

- To answer this question, the χ^2 test can be used with following hypotheses:

H_0: The considered model is correct.

H_a: The considered model is not correct.

  ▷ Under H_0, we expect $Q \approx 0$.

  ▷ Under H_a, we expect Q to be sufficiently large.

- The sampling distribution of Q under H_0 is approximately χ^2 distribution with number of degrees of freedom

  $\nu =$ number of cells (nc) $-$ number of independent parameters fitted (np) $- 1$

  Our example: nc $= 16$; np $= 1$ (i.e., $\hat{\lambda}$ from the Poisson distribution)

- To calculate the sample value (q) of Q, we need to obtain the expected counts, originating from the considered model (Poisson here).

  Our example:

  ▷ Under the hypothesized (Poisson) model, the probability that random count falls in any of the cells can be calculated from

  $$P(X = k) = \pi_k = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$

  where $\hat{\lambda} = 8.392$ (average).

  ▷ Therefore, the probability that an observation falls in the first cell (i.e., 0, 1 or 2 counts) is

  $$p_1 = \pi_0 + \pi_1 + \pi_2$$

  Furthermore, $p_2 = \pi_3$ and $p_{16} = \sum_{k=17}^{\infty} \pi_k$.

- ▷ Under assumption that $X_1, \ldots, X_{1207}$ are independent Poisson, the number of observations out of 1207 falling in given cell k follow a binomial distribution with mean $1207 \cdot p_k$

- ▷ The joint distribution of counts in all cells is multinomial with $n = 1207$ and probabilities $p_1, \ldots, p_{16}$.

  ⇒ Calculate expected number of counts in each cell
  (e.g., for cell 4: $1207 \cdot 0.0786 = 94.9$).

Result:

n	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
0-2	18	**12.2**	**2.76**
3	28	**27.0**	**0.04**
4	56	**56.5**	**0.01**
5	105	**94.9**	**1.07**
6	126	**132.7**	**0.34**
7	146	**159.1**	**1.08**
8	164	**166.9**	**0.05**
9	161	**155.6**	**0.19**
10	123	**130.6**	**0.44**
11	101	**99.7**	**0.02**
12	74	**69.7**	**0.27**
13	53	**45.0**	**1.42**
14	23	**27.0**	**0.59**
15	15	**15.1**	**0.00**
16	9	**7.9**	**0.57**
17+	5	**7.1**	**0.57**

- **Conclusion:** $P(Q > q = 8.99|H_0\ true) = 0.83 > 0.05 = \alpha.$ $\Rightarrow$ There is no substantial evidence against the Poisson distribution.

4.2 Dependency between discrete variables

- Consider the following **observed contingency table** of 2 discrete variables for a given sample:

		Variable 2				
	i/j	1	2	$\dots$	J	Total
	1	O_{11}	O_{12}	$\dots$	O_{1J}	$O_{1.}$
Variable 1	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	I	O_{I1}	O_{I2}	$\dots$	O_{IJ}	$O_{I.}$
	Total	$O_{.1}$	$O_{.2}$	$\dots$	$O_{.J}$	n

- **Research question:** Is there dependency between both variables?

- To answer this question, the χ^2 test can (again) be used with following hypotheses:

$$\boxed{\begin{array}{l} H_0\text{: Both variables are independent.} \\[1ex] H_a\text{: Both variables are dependent.} \end{array}}$$

- When H_0 is true and n is large, then Q will be approximately χ^2_ν distributed with $\nu = IJ - (I-1) - (J-1) - 1 = (I-1)(J-1)$

- The expected counts under H_0 are estimated as follows:

  ▷ $E_{ij} = n \cdot \pi_{ij}$ with unknown probability $\hat{\pi}_{ij}$ that an individual is in class corresponding to ith row and jth column.

  ▷ Under H_0: $\pi_{ij} = \pi_{i.} \cdot \pi_{.j}$, where $\pi_{i.}$ is the probability that an observation will be classified in ith row and $\pi_{.j}$ is probability that an observation will be classified in jth column.

▷ The MLE $\hat{\pi}_{i.}$ is simply the proportion of observations in the sample that are classified in the ith row, i.e., $\hat{p}_{i.} = O_{i.}/n$. Similarly: $\hat{p}_{.j} = O_{.j}/n$.

▷ $E_{ij} \approx n \cdot \hat{p}_{i.} \cdot \hat{p}_{.j} = \frac{O_{i.} \cdot O_{.j}}{n}$, leading to the following **expected contingency table**:

		Variable 2				
	i/j	1	2	$\dots$	J	Total
	1	E_{11}	E_{12}	$\dots$	E_{1J}	$E_{1.}$
Variable 1	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	I	E_{I1}	E_{I2}	$\dots$	E_{IJ}	$E_{I.}$
	Total	$E_{.1}$	$E_{.2}$	$\dots$	$E_{.J}$	n

- **Remark:** When some cells in table with expected frequencies < 5
  $\Rightarrow$ Regroup cells!!

Example:

- Consider the following **observed contingency table**:

Income

	1	2	3	4	Total
east	14	23	6	1	44
west	2	5	6	3	16
Total	16	28	12	4	60

- **Hypotheses:**:

H_0: There is independency between the income of persons and location.

H_a: There is dependency between the income of persons and location.

- The **expected contingency table** is given as follows:

Income

	1	2	3	4	Total
east	11.7	20.5	8.8	2.9	44
west	4.3	7.5	3.2	1.1	16
Total	16	28	12	4	60

- **Sample value:** $q = 10.8$.

- **Conclusion**: $P(Q \geq q) = P(Q \geq 10.8) < 0.05 = \alpha$, where $Q \sim \chi^2_3$.
  $\rightarrow$ Reject H_0.

4.3 Testing for homogeneity

- Imagine that we select subjects from several different populations and that we observe a discrete random variable for each subject.

- The χ^2 test can also be used to test whether or not the distribution of that discrete random variable is the same in each population.

  Example:

  A survey of voter sentiment was conducted in four cities to compare the fraction of voters favoring political party A. Random samples of 200 voters were polled in each city, leading to the following contingency table

opinion	1	2	3	4	total
favor A	76	53	59	48	236
do not favor A	124	147	141	152	564
total	200	200	200	200	800

  Do the data present sufficient evidence to indicate that the fractions of voters favoring party A differ in the four cities?

- The following hypotheses are then constructed:

$$\boxed{\begin{array}{l} H_0 : \pi_{i1} = \ldots = \pi_{iJ} \;\; (= \pi_{i.}) \qquad \forall i = 1, \ldots, I \\ H_a : H_0 \text{ is not true.} \end{array}}$$

  ▷ I: number of categories in the discrete random variable (here 2, i.e., favor A or not).

  ▷ J: number of populations/groups (here 4, i.e., no. of cities).

  ▷ **Our example:** π_{ij} = the fraction of voters in category i and city j.

- If we denote that fraction of voters favoring A as $\pi_{1.}$ and hypothesize that $\pi_{1.}$ is same for all cities, then first-row probabilities are all equal to $\pi_{1.}$ and the MLE is $\hat{\pi}_{1.} = 236/800$.

- The expected number of individuals favoring A equals $200 \cdot p$, which is estimated by $200 \cdot 236/800$.

- **Conclusion**: $P(Q \geq q) = P(Q \geq 10.72) = 0.01334 < 0.05 = \alpha$, where $Q \sim \chi^2_3$.
  $\rightarrow$ Reject H_0.