A. Giovanidis 2019

3. Confidence Intervals / Hypothesis Tests

Data Analysis for Networks - DataNets'19
Anastasios Giovanidis

Sorbonne-LIP6

October 03, 2019

Bibliography

A. Giovanidis 2019

B.1 H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

☞ Chapter 8.3, 8.4

A. Giovanidis 2019

Intro

We will discuss in this course two main themes:

- **Confidence Intervals**
- **Hypothesis Tests**

Applications

- Anomaly detection: Sensors observe the network ingress traffic periodically. When the network is healthy, the mean flow rate is R $[bits/sec]$. How can one decide both fast and correctly that an anomaly appears?
- Signal detection: An RF antenna needs to decide the presence or not of a signal (e.g. radar detects target)

A. Giovanidis 2019

Confidence Intervals

Interval Estimation

A. Giovanidis 2019

- Let $X_1, \ldots, X_n$ be a random sample from a distribution, with a parameter θ to be estimated.
- We have observed $x_1, \ldots, x_n$.
- We can use $\hat{\Theta} = h(X_1, \ldots, X_n)$ to estimate θ.
- Although $\hat{\Theta}$ can be asymptotically consistent, we don't know how close we are to the real θ.

☞ Introducing interval estimation: instead of giving just one estimate value $\hat{\theta}$, we produce an interval that is likely to include the true value of θ.

$$\hat{\theta} \quad \in \quad \left[\hat{\theta}_\ell,\ \hat{\theta}_h\right].$$

e.g. instead of saying $\hat{\theta} = 34.25$, we report the interval $[30.96,\ 37.81]$.

Confidence Intervals

A. Giovanidis 2019

There are two important concepts, related:

- the length of the reported interval $\hat{\theta}_h - \hat{\theta}_\ell$.
- the level of confidence about the interval.

☞ The smaller the interval, the higher the precision we estimate θ .
☞The confidence level is the probability that the constructed interval includes the real value of θ. High confidence levels are desirable.

General framework

A. Giovanidis 2019

An interval estimator with confidence level $1 - \alpha$ consists of two estimators $\hat{\Theta}_\ell\,(X_1, \ldots, X_n)$ and $\hat{\Theta}_h\,(X_1, \ldots, X_n)$ such that

$$P\left(\hat{\Theta}_\ell \leq \theta \leq \hat{\Theta}_h\right) \quad \geq \quad 1 - \alpha,$$

for every possible value of θ. Equivalently, we say that $\left[\hat{\Theta}_\ell,\ \hat{\Theta}_h\right]$ is a $(1-\alpha)100\%$ confidence interval for θ.

☞ The randomness is due to $\hat{\Theta}_\ell\,(X_1, \ldots, X_n)$ and $\hat{\Theta}_h\,(X_1, \ldots, X_n)$ and not θ.

Finding estimators

A. Giovanidis 2019

Let X be a continuous random variable with CDF $F_X(x) = P(X \leq x)$.
How can we find x_ℓ and x_h such that

$$P(x_\ell \leq X \leq x_h) \quad = \quad 1 - \alpha.$$

☞ Choose x_ℓ and x_h such that:

$$P(X \leq x_\ell) = \frac{\alpha}{2}, \quad and \quad P(X \geq x_h) = \frac{\alpha}{2}.$$

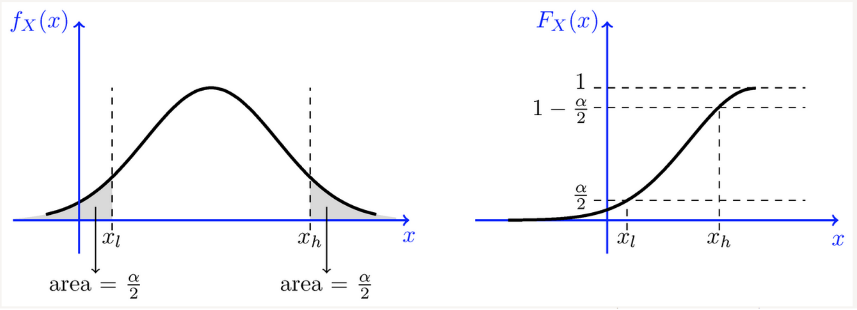
✍ This can be re-written as:

$$x_\ell = F_X^{-1}\left(\frac{\alpha}{2}\right), \quad and \quad x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Then $[x_\ell,\ x_h]$ is a $(1-\alpha)$ interval for X.

CDF and PDF

A. Giovanidis 2019



Special case: Normal r.v.

A. Giovanidis 2019

Let $Z \sim N(0,1)$, find x_ℓ and x_h such that

$$P(x_\ell \leq Z \leq x_h) \quad = \quad 0.95.$$

As we showed above,

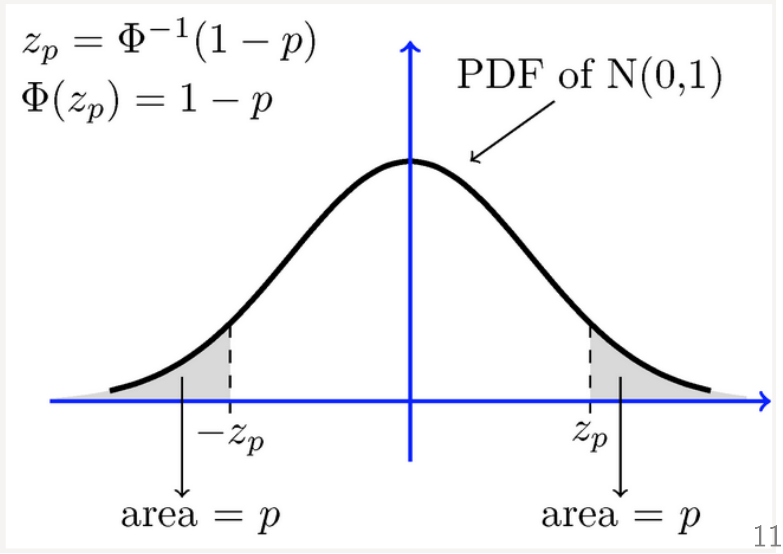
$$x_\ell = \Phi^{-1}\left(\frac{0.05}{2}\right) = -1.96, \quad \textit{and} \quad x_h = \Phi^{-1}\left(1 - \frac{0.05}{2}\right) = +1.96.$$

☞ For the Normal distribution, we denote these values by $z_{\frac{\alpha}{2}} := x_h$ and $z_{1-\frac{\alpha}{2}} := x_\ell$, and we can easily see that $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$, so that

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) \quad = \quad 1 - \alpha.$$

Normal interval

A. Giovanidis 2019



Sample Mean (from Normal)

A. Giovanidis 2019

Let $(X_1, \ldots, X_n)$ be a random sample of size n from a normal distribution $N(\theta, 1)$. Find a 95% confidence interval for θ.

$$\hat{\Theta} = \overline{X} \quad = \quad \frac{X_1 + \ldots + X_n}{n}.$$

Since $X_i \sim N(\theta, 1)$ and the X_is are i.i.d., we conclude that $\overline{X} \sim N\left(\theta, \frac{1}{n}\right)$. By normalising $\overline{X}$, we conclude that the new random variable

$$\frac{\overline{X} - \theta}{\frac{1}{\sqrt{n}}} \sim N(0, 1).$$

Note here that the above probability distribution does **not** depend on θ! We call the above random variable, a pivotal quantity. Therefore,

$$P\left(\overline{X} - \frac{1.96}{\sqrt{n}} \leq \theta \leq \overline{X} + \frac{1.96}{\sqrt{n}}\right) \quad = \quad 0.95.$$

Sample Mean (known variance)

A. Giovanidis 2019

✎ Question: Let $(X_1, \ldots, X_n)$ be a random sample of size n from a distribution with known $Var(X_i) = \sigma^2$, and unknown mean $\mathbb{E}[X_i] = \theta$. Find a $1 - \alpha$ confidence interval for θ. Assume n large.

A. Giovanidis 2019

Answer Sample Mean (known variance)

- We choose as point estimator the sample mean:

$$\hat{\Theta} = \overline{X} \quad = \quad \frac{X_1 + \ldots + X_n}{n}.$$

Since n is large, and the samples are i.i.d, we can apply the Central Limit Theorem (CLT) and conclude that

$$Q := \frac{\overline{X} - \theta}{\frac{\sigma}{\sqrt{n}}}$$

approximately follows $N(0, 1)$. Again, Q is a function of the sample and the θ, and its distribution does not depend on θ (pivotal quantity).

$$P\left(-z_{\frac{\alpha}{2}} \leq Q \leq +z_{\frac{\alpha}{2}}\right) \quad = \quad 1 - \alpha.$$

Then, $\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$ is $(1-\alpha)100\%$ confidence interval for θ.

Exercise

A. Giovanidis 2019

✎ Exercise: We wish to measure a quantity θ, but there is a random error in each measurement (noise). We take n measurements $(X_1, \ldots, X_n)$ and report the average of the measurements as the estimated value of θ. Then, measurement i is

$$X_i \quad = \quad \theta + W_i,$$

W_i being the error in the i-th measurement and all W_is are i.i.d, with $\mathbb{E}[W_i] = 0$ and $Var(W_i) = 4$ [units].

Q: How many measurements n do we need to make until we are 90% sure that the final estimation error is less than 0.25 units?

$$P(\overline{X} - 0.25 \leq \theta \leq \overline{X} + 0.25) \quad \geq \quad 0.90.$$

Solve Exercise

A. Giovanidis 2019

☞ We will use the estimator $\overline{X}$ for θ, because $\mathbb{E}[X_i] = \theta$.

• We know than the CLT applies for large n. From the above analysis, we have the formula, for a $(1-\alpha)100\%$ confidence interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

Then we have the equality

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad = \quad 0.25.$$

Here, $\alpha = 0.1$, $\sigma = \sqrt{4} = 2$, so that

$$n \quad = \quad (2z_{0.05}/0.25)^2 = \left(8 \cdot \Phi^{-1}(0.95)\right)^2 = (8 \cdot 1.645)^2 \approx 174 \textit{ samples}.$$

Sample Mean (unknown variance)

A. Giovanidis 2019

✎ Question: Let $(X_1, \ldots, X_n)$ be a random sample of size n from a distribution with unknown $Var(X_i) = \sigma^2$, and unknown mean $\mathbb{E}[X_i] = \theta$. Find a $1 - \alpha$ confidence interval for θ. Assume n large.

☞ We can not use the above discussion, because we do not know σ!

Sample Mean (unknown variance)

A. Giovanidis 2019

Two approaches:

1. Use an upper bound for σ, so that $\sigma \leq \sigma_{\max}$, (larger interval)

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma_{\max}}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma_{\max}}{\sqrt{n}}\right].$$

2. Use an estimate $\hat{\sigma}$ for σ, and we get

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right],$$

which should be relatively good for n large.

Exercise (Voters' polling)

A. Giovanidis 2019

✎ Exercise: We wish to estimate the percentage of voters that will vote for a certain candidate A in the coming elections. Let the sample size be n (large) and the unknown percentage of supporters θ.

We randomly select (with replacement) a voter and mark $X_i = 1$ if she will vote in favour of candidate A, otherwise $X_i = 0$. $X_i \sim \mathrm{Bernoulli}(\theta)$.

Q1: Find a $(1-\alpha)100\%$ confidence interval for θ based on $X_1, \dots, X_n$.

Q2: Estimate θ such that the margin of error is 3%. Assume a 95% confidence level. That is, we would like to choose n such that

$$P\left(\overline{X} - 0.03 \leq \theta \leq \overline{X} + 0.03\right) \quad \geq \quad 0.95,$$

where $\overline{X}$ is the portion of people in our random sample that say they plan to vote for the Candidate. How large does n need to be?

Answer Q1

A. Giovanidis 2019

- Since n large, we assume that the CLT holds. The interval is given by:

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma_{\max}}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma_{\max}}{\sqrt{n}}\right],$$

and we need to find an appropriate $\sigma_{\max}$.

Note, however, that

$$Var(X_i) = \theta(1-\theta) \leq \frac{1}{4} \quad \Rightarrow \quad \sigma_{\max} = \frac{1}{2}.$$

If the real θ is too small, or too large, this interval is very conservative.

Answer Q1

A. Giovanidis 2019

- • Alternatively, we can now use that

$$Var(X_i) = \theta(1-\theta) \quad \Rightarrow \quad \hat{\sigma} \stackrel{estim.}{=} \sqrt{\overline{X}(1-\overline{X})}.$$

This estimate can be replaced in the expression for the interval.

Answer Q2

A. Giovanidis 2019

- Using the expression with $\sigma_{\max}$ we get

$$z_{0.025}\frac{1}{2\sqrt{n}} = 0.03 \quad \Rightarrow \quad n = \left(\frac{1.96}{0.06}\right)^2 \approx 1068.$$

This is why most polls need a sample size of $n \approx 1000$.

A. Giovanidis 2019

Sample Mean (general)

In the most general case, when the variance is unknown, in order to find the confidence interval of the sample mean, we can use the sample standard deviation! (remember?)

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right)},$$

then the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right],$$

is approximately a $(1-\alpha)100\%$ confidence interval for θ.

✎ Exercise: If $n = 100$, $\overline{X} = 15.6$, $S^2 = 8.4$, find an approximate 99% interval for $\theta = \mathbb{E}[X_i]$.

A. Giovanidis 2019

Hypothesis Testing

Intro

A. Giovanidis 2019

☞ We need to decide whether some hypothesis is true or not.

RADAR

H_0: No aircraft is present.

H_1: An aircraft is present.

H_0 is the null hypothesis (default to be true), and
H_1 is the alternative hypothesis.

Is the coin fair?

A. Giovanidis 2019

Let θ be the probability of heads $\theta = P(HEADS)$.

H_0: The coin is fair, i.e. $\theta = \theta_0 = \frac{1}{2}$ (simple hypothesis).

H_1: The coin is not fair, i.e. $\theta \neq \frac{1}{2}$ (composite hypothesis).

☞ Given some measurements (sequential coin tosses) how can we decide whether the coin is fair or not? Let $n = 100$ tosses. Then,

$$X \sim Binomial(100, \theta)$$

is the total number of heads in the 100 tosses.

Decision criterion for coin

A. Giovanidis 2019

☞ If X is around 50 we accept H_0, because $X/100$ is close to $1/2$.

Let's use the notion of a threshold (for now unknown).

If $|X - 50| \leq t$, accept H_0,

if $|X - 50| > t$, accept H_1.

Some issues here:
- What should be the value of t?
- What guarantees does the choice of t offer to our decision?

Error Probability Types

A. Giovanidis 2019

Type I Error (False Positive)

$$P(type\ I\ error) = P\left(accept\ H_1 \mid H_0\right) \leq \alpha$$

Type II Error (False Negative)

$$P(type\ II\ error) = P\left(accept\ H_0 \mid H_1\right) \leq \beta$$

α, β are the levels of significance.

☞ Exercise: Calculate the threshold in the coin-toss example, so that

$$P(type\ I\ error) = P\left(|X - 50| > t \mid H_0\right) = \alpha = 0.05$$

Solve Exercise

A. Giovanidis 2019

Use the CLT to approximate the distribution of the mean by the $\mathcal{N}(0,1)$:

$$Y = \frac{\frac{X}{n} - \theta_0}{\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{n}}} = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = \frac{X-50}{5}$$

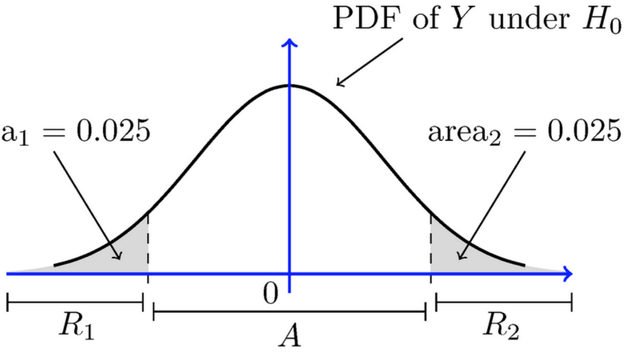
$$\begin{aligned} P(type\ I\ error) = P\left(|X-50| > t \mid H_0\right) &= P\left(\left|\frac{X-50}{5}\right| > \frac{t}{5} \mid H_0\right) \\ &= P\left(|Y| > \frac{t}{5} \mid H_0\right) = 0.05. \end{aligned}$$

We can write (due to symmetry of the Normal):

$$\begin{aligned} 2 \cdot P(Y > \frac{t}{5} \mid H_0) &= 2 - 2\Phi\left(t/5\right) = 0.05 \\ \Rightarrow t &= 5\Phi^{-1}(0.975) = 5 \cdot 1.96 = 9.8 \end{aligned}$$

A. Giovanidis 2019

Accept and Reject region



A = Acceptance Region

$R = R_1 \cup R_2$ = Rejection Region

$\alpha = \text{P(type I error)} = \text{area}_1 + \text{area}_2 = 0.05$

Solve Exercise cont'd

A. Giovanidis 2019

The decision criterion for the coin becomes:

If $40.2 \leq X \leq 59.8$, accept H_0,

if $X > 59.8$ or $X < 40.2$, accept H_1.

To present this result better:

- The acceptance region is $A = \{41, 42, \ldots, 59\}$.
- The rejection region is $R = \{0, \ldots, 40\} \cup \{60, \ldots, 100\}$

Test Statistic

A. Giovanidis 2019

DEFINITION
Let $X_1, X_2, \ldots, X_n$ be a random sample of interest. A statistic is a real-valued function of the data. e.g. the sample mean,

$$W(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

is a statistic. A test statistic is a statistic based on which we build our test.

In the above example the statistic was X and :

$$Y = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}.$$

Exercise Radar

A. Giovanidis 2019

RADAR

H_0: No aircraft is present.

H_1: An aircraft is present.

The received signal is

$$X = \theta + W = \begin{cases} W, & \text{no aircraft} \\ 1 + W, & \text{if aircraft is present} \end{cases}$$

where $\theta \in \{0, 1\}$ and $W \sim \mathcal{N}\left(0, \sigma^2 = 1/9\right)$.

Solution Radar

A. Giovanidis 2019

Under H_0: $X \sim \mathcal{N}(0, 1/9)$, while, under H_1: $X \sim \mathcal{N}(1, 1/9)$.

RADAR Decision. Choose a threshold c

If $X \leq c$: accept H_0.

If $X > c$: accept H_1.

$$
\begin{aligned}
P(\mathit{type\ I\ error}) &= P\left(\mathit{Reject}\ H_0 \middle|\ H_0\right) = P(X > c \mid H_0) \\
&= P\left(W > c\right) = 1 - \Phi(3c)
\end{aligned}
$$

Letting $\alpha = 0.05$ we obtain

$$
c = \frac{1}{3}\Phi^{-1}(1-\alpha) = \frac{1}{3}\Phi^{-1}(1-0.05) = 0.548
$$

Solution Radar cont'd

A. Giovanidis 2019

$$
\begin{aligned}
P(type\ II\ error) &= P\left(Accept\ H_0 |\ H_1\right) = P(X \leq c \mid H_1) \\
&= P\left(1 + W \leq c\right) = \Phi(3(c-1))
\end{aligned}
$$

Since we found $c = 0.548$ we get

$$
\beta = \Phi(-1.356) = 0.088.
$$

Solution Radar cont'd II

A. Giovanidis 2019

If we want to choose $\alpha = 0.01$, then the threshold takes the value:

$$c = \frac{1}{3}\Phi^{-1}(1-\alpha) = \frac{1}{3}\Phi^{-1}(1-0.01) = 0.775$$

☞ Suppose we measure $X = 0.6$. Then for $\alpha = 0.05$ we get $0.6 > 0.548$ and we reject H_0 (an airplane is detected). But, for $\alpha = 0.01$ we get $0.6 < 0.775$ and we accept H_0 (no airplane).

Solution Radar cont'd III

A. Giovanidis 2019

If we want the probability of missing a present aircraft to be less than 5%,

$$\begin{aligned} 0.05 &= \Phi(3(c-1)) \Rightarrow \\ c &= 1 + \frac{1}{3}\Phi^{-1}(0.05) = 0.452 \end{aligned}$$

☞ Thus for type I error significance level, we get:

$$\alpha = 1 - \Phi(3c) = 1 - \Phi(3 \cdot 0.452) = 0.0875.$$

Trade-off between α and β

A. Giovanidis 2019

☞ We cannot make both α and β small simultaneously: trade-off.

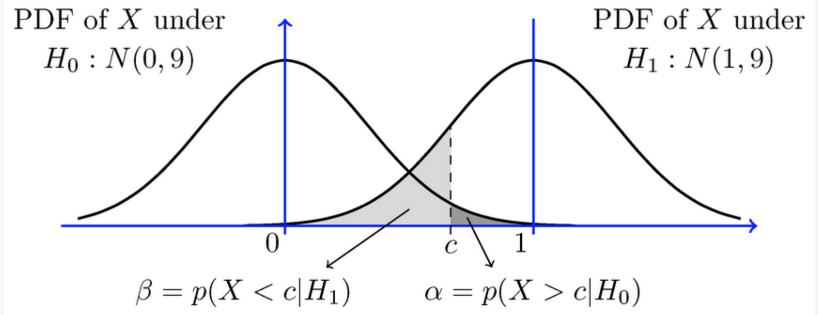
Take a look at the RADAR example:

$$\begin{aligned} \alpha &= 1 - \Phi(3c) \\ \beta &= \Phi(3(c-1)) \end{aligned}$$

Since $\Phi(y)$ is increasing with y, we see that when the c threshold increases, then α decreases, but β increases!

Trade-off cont'd

A. Giovanidis 2019



Test for the mean

A. Giovanidis 2019

Consider a random sample $X_1, \ldots, X_n$ from a distribution, and make an inference for the mean

H_0: $\mu = \mu_0$,

H_1: $\mu \neq \mu_0$.

☞: Test statistic is the normalised sample mean

$$W(X_1, \ldots, X_n) \quad = \quad \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

and use S instead of σ if the standard deviation is unknown.

Threshold for the mean

A. Giovanidis 2019

If we assume H_0, then the test statistic $W \sim \mathcal{N}(0,1)$.
☞ Choose a threshold, c:
If $|W| \leq c$ we accept H_0 and if $|W| > c$ we accept H_1.
Type I error:

$$\alpha = P(|W| > c \mid H_0) = 2 \cdot P(W > c \mid H_0).$$

Thus, we conclude $P(W > c \mid H_0) = \alpha/2$.
Therefore,

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = z_{\frac{\alpha}{2}}$$

and we accept H_0 if $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}$, and we reject it otherwise.

Relation to confidence intervals

A. Giovanidis 2019

The condition to accept the H_0 for the mean μ_0

$$\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \quad \leq \quad z_{\frac{\alpha}{2}}$$

can be rewritten as

$$\mu_0 \quad \in \quad \left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

☞ This is the $(1-\alpha)100\%$ confidence interval for μ_0. (slide 14/43)

P-values

A. Giovanidis 2019

> The P-value is the lowest significance level α that results in rejecting the null hypothesis.

Given a threshold c we reject the H_0 if the test statistic has a value larger than the threshold. The smaller the required significance level, the larger the threshold.

If the P-value is small, then the threshold is quite high, and the test statistic even higher, so it is very unlikely to have occurred under H_0, and we are more confident in rejecting the null hypothesis.

☞How do we find P-values? Let's look at an example.

Finding P-values

A. Giovanidis 2019

The H_0 hypothesis is rejected when $W > c$ for the test statistic.
Let w be the realisation of the test statistic W of a given sample.

> The P-value is the type I error for $c = w$

Example: coin toss. Let us use $W = \frac{X-50}{5}$, so for $X = 60$, $w = 2$

$$\begin{aligned} P - value &= P(type\ I\ error\ for\ c = 2) \\ &= P\left(W > 2\right) = 1 - \Phi(2) = 0.023. \end{aligned}$$

Likelihood Ratio Tests (LRT)

A. Giovanidis 2019

Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. The likelihood function is defined for discrete and continuous variables:

$$L(x_1, \ldots, x_n; \theta) = P_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$$
$$L(x_1, \ldots, x_n; \theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$$

To decide between two hypotheses:
$H_0: \ \theta = \theta_0$ and $H_1: \ \theta = \theta_1$
we define the likelihood ratio test

$$\lambda(x_1, \ldots, x_n) = \frac{L(x_, \ldots, x_n; \theta_0)}{L(x_, \ldots, x_n; \theta_1)},$$

and we decide for H_0 if $\lambda \geq c$ else for H_1 if $\lambda < c$. The c is chosen based on the desired α.

Exercise LRT

A. Giovanidis 2019

☞ Exercise: We look again at the RADAR problem. We observe the random variable:

$$X = \theta + W,$$

where $W \sim \mathcal{N}\left(0, \sigma^2 = 1/9\right)$. We need to decide between
$H_0 : \theta = \theta_0 = 0$,
$H_1 : \theta = \theta_1 = 1$.
Let a single observation $X = x$.

Design a level $\alpha = 0.05$ test to decide between H_0 and H_1.

A. Giovanidis 2019

END