CS145: Probability & Computing

Lecture 19: Hypothesis Testing

Instructor: Cyrus Cousins

Brown University Computer Science

Figure credits:
Bertsekas & Tsitsiklis, ***Introduction to Probability****, 2008*
Pitman, ***Probability****, 1999*
Credits:
Hui Wang, ***Hypothesis Testing****, 2017*

From Probability to Statistics

- In *probability theory* we compute the probability that 20 independent flips of a fair (unbiased) coin give the sequence

  HTTHTHTHHTTHTHTHHTTT

- In *statistics* we ask: Given that we observed the sequence

  HTTHTHTHHTTHTHTHHTTT

  what is the likelihood that the coin is fair (unbiased)?

https://hattonsoflondon.co.uk/

Hypothesis Testing and Coinflips

Over Spring 2009 two Berkeley undergraduates, Priscilla Ku and Janet Larwood, undertook a task to perform 40,000 coin tosses.

It was "only" one hour per day for a semester....

Result:

Heads = 20217 times.

Tails = 19783 times.

$$\hat{p} = \frac{20217}{40000}$$

Question: Is the coin fair?

Hypothesis Testing Steps

- Formulate your theory “in a testable way”
  - Null Hypothesis
  - Alternative Hypothesis
- Identify your test
  - Test statistics
- Identify how certain you want to be
  - Level of Significance
- Decision criteria
  - Identify a “rejection” region
  - p-value

What Is a Hypothesis

- A hypothesis is a claim (assumption) about a population parameter:

  - population mean

    Example: The mean monthly cell phone bill of this city is $\mu = \$42$

  - population proportion

    Example: The proportion of adults in this city with cell phones is $p = 68$

The Null Hypothesis H_0

Usually refers to the default position

- New theory does not give better explanation
- New medication is not performing better
- …

Hypothesis testing is not symmetric. It gives priority to the null.

The null hypothesis is rejected only if the data shows that it is very unlikely, otherwise the null holds.

The Null Hypothesis H_0

States the assumption (numerical) to be tested

Example: the coin is fair $H_0 : p = 0.5$

where p is the probability of head

Is always about a population (or data distribution) parameter, NOT about a sample statistic

$H_0 : p = 0.5$ ✓

$H_0 : \hat{p} = 0.5$ ✗

Null Hypothesis

We need to decide regarding our coin…

- H_1: The alternative hypothesis - the coin is weighted
- H_0: The null hypothesis - the coin is fair

Researchers do not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Basic Frequentist Idea

- Hypotheses are fixed: they synthesize a prior belief on the data

- Data is random: the analyst evaluates if the hypothesis is coherent with respect to the random data

The Hypothesis Testing Setup

1. Set Up *Null Hypothesis* (H_0) and *Alternative Hypothesis* (H_a):

$$H_0 : p = 0.5 \qquad H_\mathrm{a} : p \neq 0.5$$

Null hypothesis is the claim to be tested.

Hypothesis testing evaluates the strength of empirical evidence against null hypothesis.

2. Finding a Test Statistic: What else but empirical frequency $\hat{p}$.

3. Determine If Data Is Plausible, assuming null hypothesis is correct

The Hypothesis Testing Setup

1. Set Up *Null Hypothesis* (H_0) and *Alternative Hypothesis* (H_a):

$$H_0 : p = 0.5 \qquad H_{\mathrm{a}} : p \neq 0.5$$

Null hypothesis is the claim to be tested.

Hypothesis testing evaluates the strength of empirical evidence against null hypothesis.

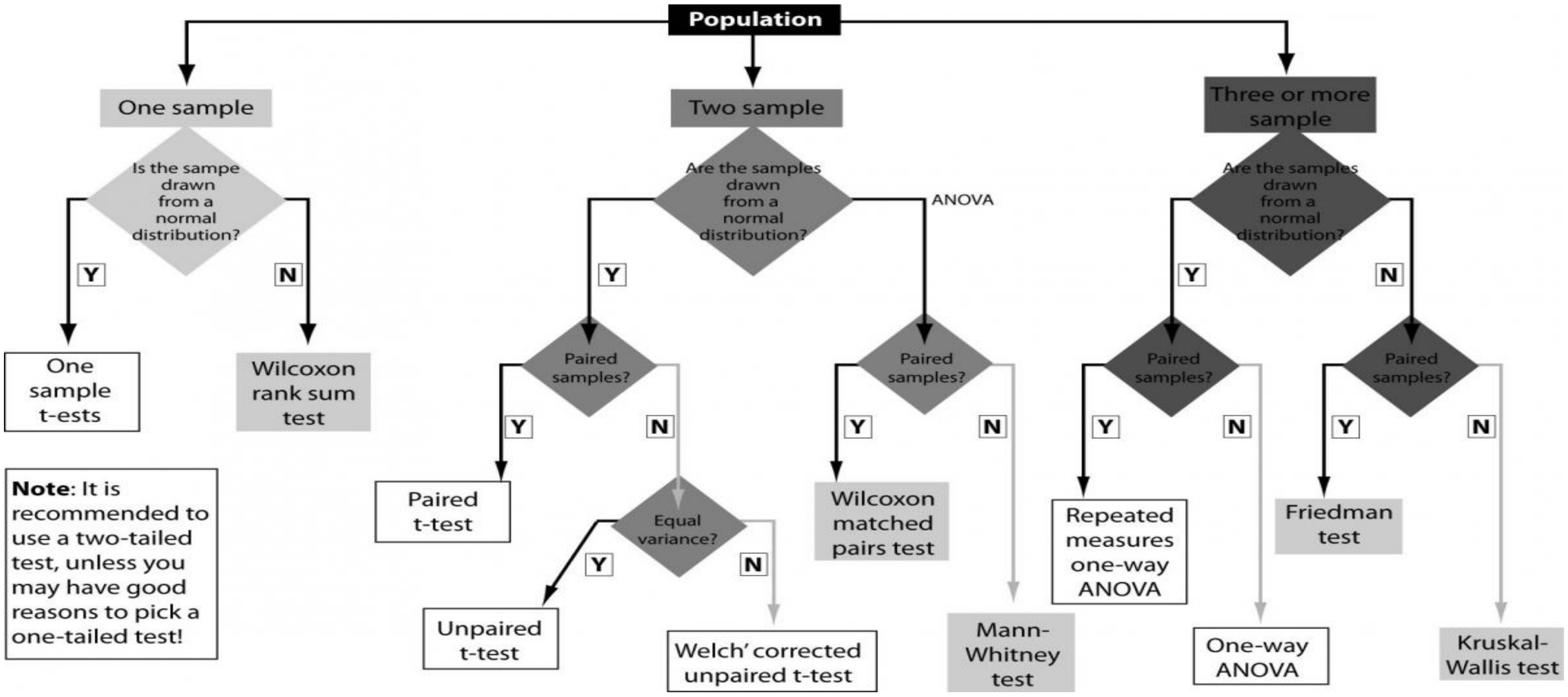
2. Finding a Test Statistic: What else but empirical frequency $\hat{p}$.

3. Determine If Data Is Plausible, assuming null hypothesis is correct
…. i.e. assuming the population distribution corresponds to H_0

Select Your Test

- Testing is a bit like finding the right recipe based on these ingredients:
  - Question
  - Data type
  - Sample size
  - Variance known? Variance of several groups equal?
- Good news: Plenty of tables available, e.g.,
  - http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm (with examples in R, SAS, Stata, SPSS)

How to Choose Your Test



Example of a Table of Tests

Summary Table for Statistical Techniques

Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
Estimating a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• What is the average weight of adults? • What is the average cholesterol level of adult females?	1-sample t-interval $\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately norr or • have a large sample size ($n \geq 30$)
Test about a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• Is the average GPA of juniors at Penn State higher than 3.0? • Is the average Winter temperature in State College less than 42° F?	$H_o: \mu = \mu_o$ $H_a: \mu \neq \mu_o$ or $H_a: \mu > \mu_o$ or $H_a: \mu < \mu_o$ The one sample t test: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately norr or • have a large sample size ($n \geq 30$)
Estimating a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• What is the proportion of males in the world? • What is the proportion of students that smoke?	1-proportion Z-interval $\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	Stat >Basic statistics >1-sample proportion	• have at least 5 in each category
Test about a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• Is the proportion of females different from 0.5? • Is the proportion of students who fail Stat 500 less than 0.1?	$H_o: \pi = \pi_o$ $H_a: \pi \neq \pi_o$ or $H_a: \pi > \pi_o$ or $H_a: \pi < \pi_o$ The one proportion Z-test: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Stat >Basic statistics >1-sample proportion	• $n\pi_o \geq 5$ and $n(1-\pi_o) \geq$

http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf

Hypothesis Testing with P-Values

What Is p-value?

It is the probability of observing test statistics that are as extreme or more extreme than the present empirical data, assuming H_0 is valid.

$$p = Pr\left(\left|\hat{p} - \frac{1}{2}\right| \geq \frac{20217}{40000} - \frac{1}{2}\right)$$
$$= Pr\left(\left|\hat{p} - \frac{1}{2}\right| \geq 0.005425\right)$$

Find p-value: Under null hypothesis H_0, $\hat{p}$ is approximately N(0.5, 0.0025). Answer: p-value = 0.03

P-value computation depends on the specific assumptions/test being used!

Hypothesis Testing with P-Values

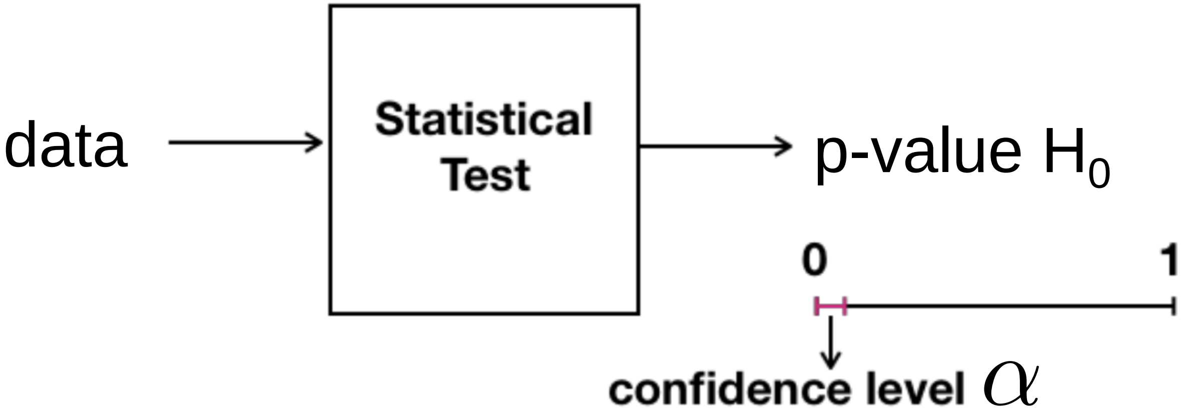
What Is Significance (Confidence) Level α?

It is an artificially given threshold such that any empirical observation that falls in the top α proportion of the most extreme scenarios (under null hypothesis H_0) is deemed implausible. The most common value of α is 5%, even though α = 1% is also widely used.

> Null hypothesis is rejected if and only if P-value is less than the significance level α.

Why is This Working?

If H_0 is a true null (i.e., should NOT be rejected) then its p-value is uniformly distributed in [0,1]



Hence, P(p-value $H_0 \leq \alpha$) $= \alpha$

Thus, P(H_0 rejected | H_0 is a true null) $= \alpha$

Comments on Hypothesis Testing

- Relation Between P-Value and Significance Level α: Null hypothesis is rejected if and only if P-value is less than the significance level α.

- Rejection and Failure to Reject: Rejection of null hypothesis does not mean null hypothesis is wrong. It means null hypothesis is statistically implausible. Similarly, failure to reject null hypothesis does not mean it is correct. It means null hypothesis is not statistically implausible.

- Statistical Significance: Statistical significance is not practical significance — recall the 40000 coin tosses. A small practical discrepancy can be statistically very significant, especially with large data set!

Types of Error

Outcome
(Probability)

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	**No error** **(1 - α)**	**Type II Error** **(β)**
Reject H_0	**Type I Error** **(α)**	**No Error** **(1 - β)**

Types of Error and the Power of Tests

- **Type I Error (False Positive)**: Given a significance level α, what is the chance that null hypothesis will be rejected, even when it is indeed correct?

  Answer: $P(H_0 \text{ rejected } | H_0 \text{ is true}) = \alpha$

- **Type II Error (False Negative)**: Given a significance level α, what is the chance that null hypothesis will fail to be rejected, even when it is indeed wrong?

  Answer: $\beta = 1 - P(H_0 \text{ is rejected} | H_1 \text{ is true})$

The ideal scenario is that both α and β are small. But they are in conflict! Everything else being equal, one cannot reduce type I error and type II error simultaneously.

- **Power of Tests:** It is defined to be $1 - \beta = P(H_0 \text{ is rejected} | H_1 \text{ is true})$

Avoiding False Positives

- Usually we are looking for sufficient evidence to reject H_0.

- Type I errors are implicitly more important than type II errors.

- One usually controls type I error below some prefixed small threshold, and then, subject to this control, look for a test which maximizes power or minimizes type II error.

Testing Means of Normals

- Let $\{X_1, \ldots, X_n\}$ be iid samples from $N(\mu, \sigma^2)$, where σ^2 is known but μ unknown. Want to perform hypothesis testing on μ.

- We consider three scenarios.

  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu < \mu_0$
  - Two-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$

Testing Means of Normals

- Let $\{X_1, . . ., X_n\}$ be iid samples from $N(\mu, \sigma^2)$, where σ^2 is known but μ unknown. Want to perform hypothesis testing on μ.

- We consider three scenarios.

  - Upper (Right) Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - Lower (Left) Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu < \mu_0$
  - Two-Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$

Testing Means of Normals: Upper Tailed Test

- Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu > \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!

   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

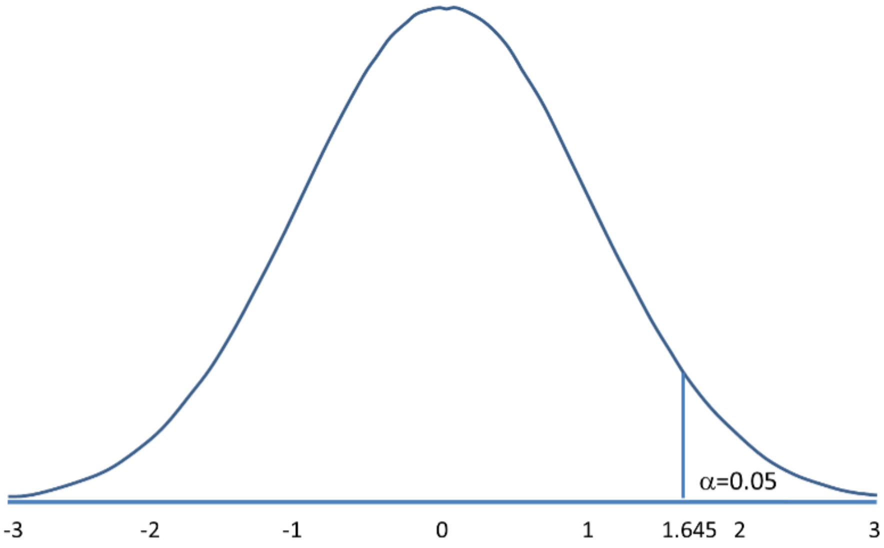
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p = P(\bar{X} \geq \bar{x}) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(-\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α , we reject H_0 iff $\alpha \geq p - value$

Upper (Right) Tailed Test



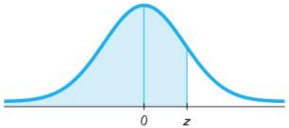
The decision rule is: Reject H_0 if $Z \geq 1.645$.

Upper-Tailed Test	
α	Z
0.10	1.282
0.05	1.645
0.025	1.960
0.010	2.326
0.005	2.576
0.001	3.090
0.0001	3.719

Use the Appropriate Table!

Use tables for the Standard Normal Distribution (z-tables)

Report the cumulative area from the LEFT

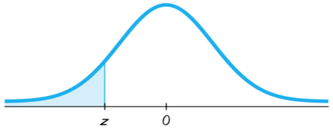


POSITIVE *z* Scores

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

EGATIVE *z* Scores



2 Standard Normal (*z*) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	.0001									
	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084

Example

NEGATIVE *z* Scores

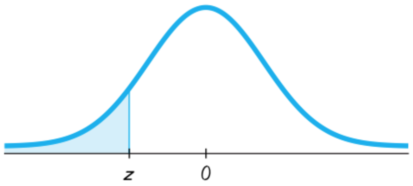


TABLE A-2 Standard Normal (*z*) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084

Suppose we need $\Phi\left(-2.45\right)$

Testing Means of Normals: Lower Tailed Test

➢ Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu < \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!

   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

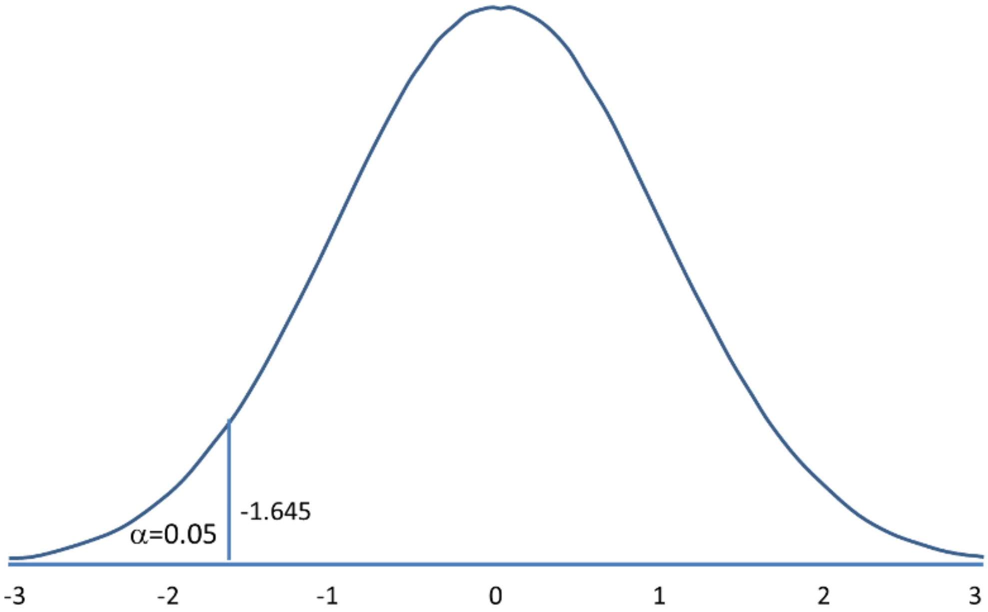
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p = P(\bar{X} \leq \bar{x}) = \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α, we reject H_0 iff $\alpha \geq p - value$

Low (Left) Tailed Test



The decision rule is: Reject H_0 if $Z \leq 1.645$.

Lower-Tailed Test	
a	Z
0.10	-1.282
0.05	-1.645
0.025	-1.960
0.010	-2.326
0.005	-2.576
0.001	-3.090
0.0001	-3.719

Testing Means of Normals: Two Sided Test

- Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu \neq \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!

   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

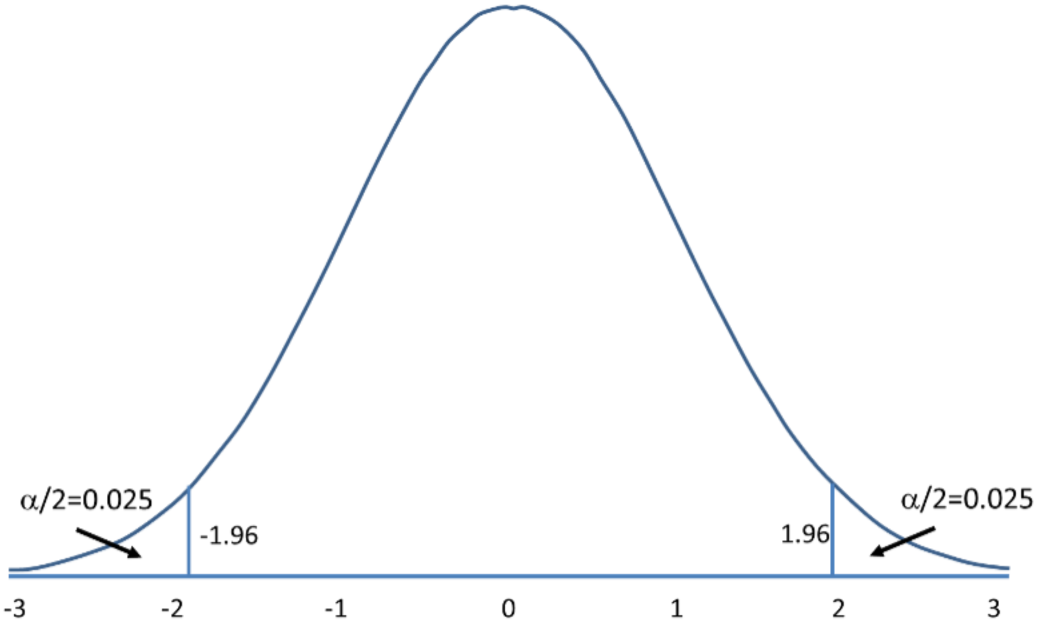
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p = P(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|) = 2\Phi\left(-\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α , we reject H_0 iff $\alpha \geq p - value$

Two Tailed Test



The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Two-Tailed Test	
α	**Z**
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

Extension to Large Samples

- The results on testing means of normals can be extended to large sample test where the test statistic is approximately (in the asymptotic sense) normally distributed.
- Common examples are given by the z-test (for >30 sample points) and the t-test (to be used with a lower number of samples)

- Example 1 – Testing mean: let $\{X_1, . . ., X_n\}$ be iid samples from some population distribution with unknown mean μ.
  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - Two-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$
- Test statistic is sample mean $\bar{X}$. By central limit theorem $\bar{X} \rightarrow N(\mu, \sigma^2/n)$

  All formulae we have obtained previously are valid.
- When σ is unknown, one can use sample standard deviation s in place of σ

Extension to Large Samples

- The results on testing means of normals can be extended to large sample test where the test statistic is approximately (in the asymptotic sense) normally distributed.
- Common examples are given by the z-test (for >30 sample points) and the t-test (to be used with a lower number of samples)

- Example 2 – Testing proportions: let $\{X_1, . . ., X_n\}$ be iid Bernoulli samples such that $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$
  - One-Sided Test: $H_0 : p = p_0$, $H_a : p > p_0$
  - Two-Sided Test: $H_0 : p = p_0$, $H_a : p \neq p_0$
- Test statistic is sample mean $\bar{X}$. By central limit $\bar{X} \to N(p, p(1-p)/n)$

  All formulae we have obtained previously are valid with p_0 in place of μ_0 and $p_0(1-p_0)$ in place of σ^2

Example: One Sample Z-Test

- Suppose we have a sample with: $\bar{x} = 0.52, \sigma = 7.89, n = 27$

$$H_0 : \mu = 0, \qquad H_a : \mu > 0$$

- Compute standard z-test statistic:

$$z = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{0.52}{7.89/\sqrt{27}} = 0.3425$$

- Compute p-value: $\Phi(-z) = \Phi(-0.3425 = 0.366)$

- Decision: for $\alpha = 0.05$ we fail to reject H_0 as $0.366 > 0.05$

Example: Two Sample Z-Test

- Compare two population means: Do indoor cats live longer than outdoor ones?

Cats	Sample size	Mean age	Sample Std
Indoor	64	14	4
Wild	36	10	5

- State hypotheses: let μ_I (resp., μ_O) denote the true population mean age of indoor (resp., outdoor) cats

$$H_0 : \mu_I = \mu_O, \quad H_a : \mu_I > \mu_O$$

- Test statistic: difference in population means

$$\bar{d} = \bar{x}_I - \bar{x}_O = 14 - 10 = 4$$

Example: Two Sample Z-Test

- Characterize distribution $\bar{D} = \bar{X}_I - \bar{X}_O$:

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{4^2}{64} + \frac{5^2}{36}} = 0.97$$

- p-value computation: under the null hypothesis $\bar{D} \sim N(0, \sigma_{\bar{D}}^2)$

$$p = P(N(0, 0.97^2) \geq 4) = 1 - \Phi\left(\frac{4}{0.97}\right) \leq 0.00003$$

- Decisions: for confidence $\alpha = 0.05$ we reject the null hypothesis

Example: Comparing Two Proportions

- In order to test if there is any significant difference between opinions of males and females on gun ban, random samples of 100 males and 150 females were taken.

Sex	Sample size	Favor	Oppose
Male	100	52	48
Female	150	95	55

- Set up the hypotheses: let p_F (resp., p_M) be the fraction of females (resp., males) which support gun ban.

$$H_0 : p_F = p_M, \qquad H_a : p_F \neq p_M$$

Example: Comparing Two Proportions

- Test statistic: difference in sample (empirical) proportions:

$$\bar{d} = \bar{p}_F - \bar{p}_M = \frac{52}{100} - \frac{95}{150} = -0.113$$

- Distribution of difference of sample proportions: $\bar{D}$ is approximately normal with $\mu = p_F - p_M$ and:

$$\sigma_{\bar{D}} = \sqrt{\frac{p_F(1-p_F)}{n_F} + \frac{p_M(1-p_M)}{n_M}}$$

- Pooled estimate: Under the null hypothesis $p_F = p_M$. Hence we can compute a pooled estimate for $p_F = p_M$ as:

$$\bar{p} = \frac{52+95}{100+150} = 0.588$$

Example: Comparing Two Proportions

- p-value: Under the null-hypotheis we have $\bar{D} \sim N(0, \sigma_{\bar{D}}^2)$, where:

$$\sigma_{\bar{D}}^2 = \bar{p}(1 - \bar{p}) \left(\frac{1}{n_F} + \frac{1}{n_F} \right)$$

$$= 0.588(1 - 0.588)(100^{-1} + 150^{-1}) = 0.0040373316$$

$\sigma_{\bar{D}}^2$ is the sample variance

Two tailed test, hence

$$p = P(|\bar{D} - \mu_0| \geq |\bar{d} - \mu_0|) = 2\Phi\left(-\frac{|\bar{d} - \mu_0|}{\sigma_{\bar{D}}} \right)$$

$$= 2\Phi\left(-\frac{0.113}{0.6354} \right) = 0.075$$

Example: Comparing Two Proportions

- Decision: given the confidence level $\alpha = 0.05$, we fail to reject the null hypothesis, and, thus we reject the alternative hypothesis.

- There is no statistically significant evidence that suggests male and females have different opinions on gun ban.