University of BRISTOL

spring semester 2019

Hypothesis Testing (1) - Lecture 3

APPLIED STATISTICS - EMAT 30007

Nikolai Bode and Ksenia Shalonova

Department Of Engineering Mathematics

Parameter estimation vs hypothesis testing (1)

There are two types of questions in statistical inference.

Parameter estimation (confidence intervals): What parameter values would be consistent with the sample data?

Example (Estimating μ)

A manufacturer of muesli bars needs to describe the average fat content of the bars (the mean, μ , of the distribution of fat contents that would be produced using the recipe). Several bars are analysed and their fat contents are measured.

Parameter estimation vs hypothesis testing (2)

Hypothesis testing: Are the sample data consistent with some statement about the parameters?

Example (Testing if $\mu = 4.2$)

A particular brand of muesli bar is claimed by the manufacturer to have a fat content of 4.2 g per bar. A consumer group suspects that the manufacturer is understating the fat content, so a random sample of bars is analysed. The consumer group must assess whether the data are consistent with the statement (or hypothesis) that the underlying mean fat content for this type of bar is $\mu = 4.2$ g.

Null and Alternative Hypothesis (1)

Null Hypothesis H_0 often specifies a single value for the unknown parameter such as $'\alpha = ...'$. It is a `default` value that can be accepted as holding if there is no evidence against it. A researcher often collects data with the express hope of disapproving the null hypothesis.

If the null hypothesis is not true, we say that the alternative hypothesis H_A holds.

Are the data consistent with the null hypothesis? If the data are not consistent with the null hypothesis, then we can conclude that the alternative hypothesis must be true.

Either the null hypothesis or the alternative hypothesis must be true.

Null and Alternative Hypothesis (2)

2	3	4	5	6	7	8	9	12	13
413	90	74	55	23	97	50	359	487	102
14	10	57	320	261	51	44	9	18	209
58	60	48	65	87	11	102	12	100	14
37	186	29	104	7	4	72	270	7	57
100	61	502	220	120	141	22	603	98	54
65	49	12	239	14	18	39	3	5	32
9	14	70	47	62	142	3	104	85	67
169	24	21	246	47	68	15	2	91	59
447	56	29	176	225	77	197	438	43	134
184	20	386	182	71	80	188		230	152
36	79	59	33	246	1	79		3	27
201	84	27	15	21	16	88		130	14
118	44	153	104	42	106	46			230
34	59	26	35	20	206	5			66
31	29	326		5	82	5			61
18	118			12	54	36			34
18	25			120	31	22			
67	156			11	216	139			
57	310			3	46	210			
62	76			14	111	97			
7	26			71	39	30			
22	44			11	63	23			
34	23			14	18	13			
	62			11	191	14			
	130			16	18				
	208			90	163				
	70			1	24				
	101			16					
	208			52					
				95					

The data show the number of operating hours between successive failures of air-conditioning equipment in ten aircrafts. The sample of 199 values is a test statistic.

We can test the manufacturer's claim that the rate of failures is no more than one per 110 hours of use.

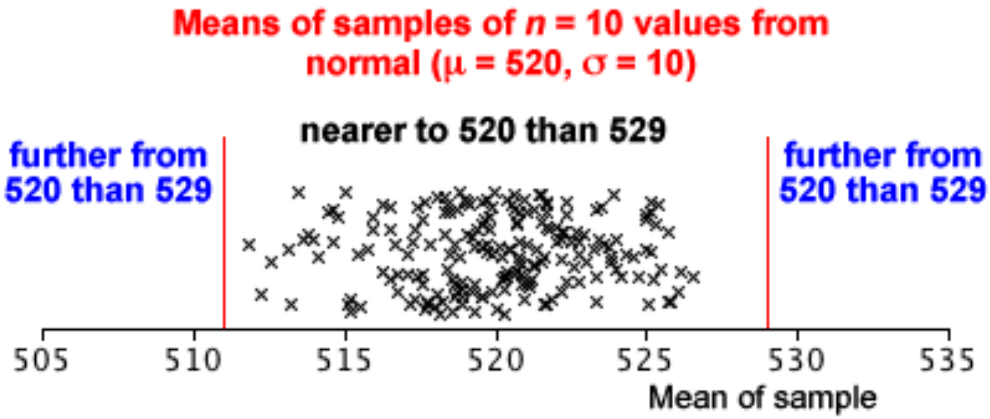
H_0: $\lambda \leq \frac{1}{110}$ (claim of a manufacturer)
H_A:$\lambda > \frac{1}{110}$

Can be simplified:

H_0: $\lambda = \frac{1}{110}$ (claim of a manufacturer)
H_A:$\lambda > \frac{1}{110}$

P-value - intuition

In an industrial process some measurement is normally distributed with standard deviation $\sigma = 10$. Its mean should be $\mu = 520$, but can differ a little bit. Samples of $n = 10$ measurements are regularly collected as part of quality control. If a sample had $\overline{(x)} = 529$, does the process need to be adjusted?

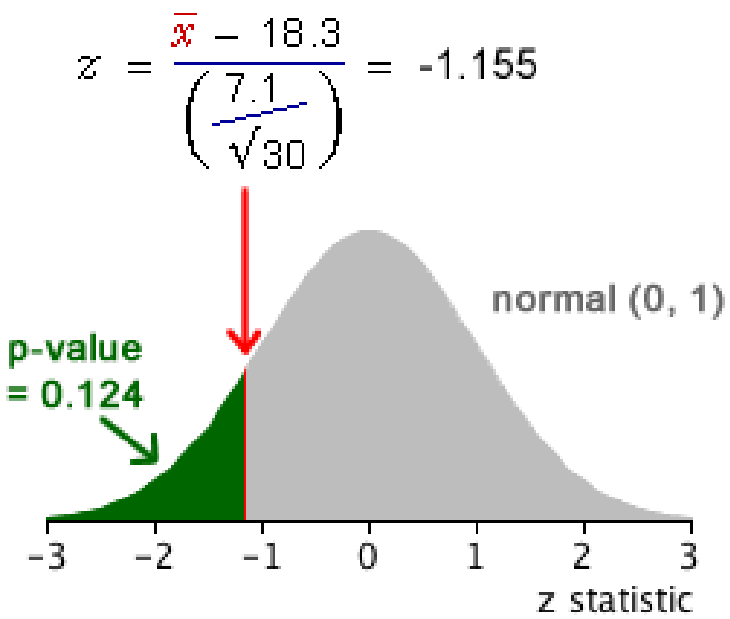


From the 200 simulated samples above (Monte Carlo simulation), it seems very unlikely that a sample mean of 529 would have been recorded if $\mu = 520$.

There is strong evidence that the industrial process no longer has a mean of $\mu = 520$ and needs to be adjusted.

P-value for mean in normal distribution with known σ - one-tailed test

We are given a random sample of $n = 30$ with $\bar{x} = 16.8$. Does the population have mean $\mu = 18.3$ and standard deviation $\sigma = 7.1$, or is the mean now lower than 18.3?



H_0: $\mu = 18.3$
H_A: $\mu < 18.3$

The p-value can be evaluated using the statistical distance of 16.8 from 18.3 (a z statistic).

Conclusion: the p-value is reasonably large, meaning that a sample mean as low as 16.8 would not be unusual if $\mu = 18.3$, so there is no evidence against H_0.

P-value for mean in normal distribution with known σ - two-tailed test

Example

Companies test their products to ensure that the amount of active ingredient is within some limits. However the chemical analysis is not precise and repeated measurements of the same specimen usually differ slightly. One type of analysis gives results that are *normally distributed* with a mean that depend on the actual product being tested and standard deviation 0.0068 grams per litre. A product is tested three times with the following concentrations of the active ingredient: 0.8403, 0.8363 and 0.8447 grams per litre. Are the data consistent with the target concentration of 0.85 grams per litre?

Null Hypothesis $H_0 : \mu = 0.85$
Alternative Hypothesis $H_A : \mu \neq 0.85$
test statistic with $\bar{X} = 0.8404$ is used: $z = \frac{0.8404 - 0.85}{0.0068/\sqrt{3}} = -2.437$ with lower tail of standard normal distribution: $P(Z \leqslant -2.437) = 0.00741$
p-value $= 2 \times 0.00741 = 0.0148$

p-value interpretation: there is moderately strong evidence that the true concentration μ

P-value for mean in normal distribution with unknown σ - one-tailed test

Example

Both cholesterol and saturated fats are often avoided by people who are trying to lose weight or reduce their blood cholesterol level. Cooking oil made from soybeans has little cholesterol and has been claimed to have only 15% saturated fat. A clinician believes that the saturated fat content is greater than 15% and randomly samples 13 bottles of soybean cooking oil for testing with the following percentage saturated fat: 15.2 12.4 15.4 13.5 15.9 17.1 16.9 14.3 19.1 18.2 15.5 16.3 20.0.

$H_0 : \mu = 15; H_A : \mu > 15$

T test for μ: $t = \frac{16.138-15}{2.154/\sqrt{13}} = 1.906$

p-value for the test is the upper tail area of the `t-distribution` with 12 degrees of freedom: $P(T \geq 1.906) = 0.040$

p-value interpretation: Since this is below 0.05, we conclude that there is moderately strong evidence that the mean saturated fat content of the oils is higher than the claimed 15%.

P-value

A p-value describes the evidence `against` H_0.

A p-value is evaluated from a random sample so it has a distribution in the same way that a sample mean has a distribution.

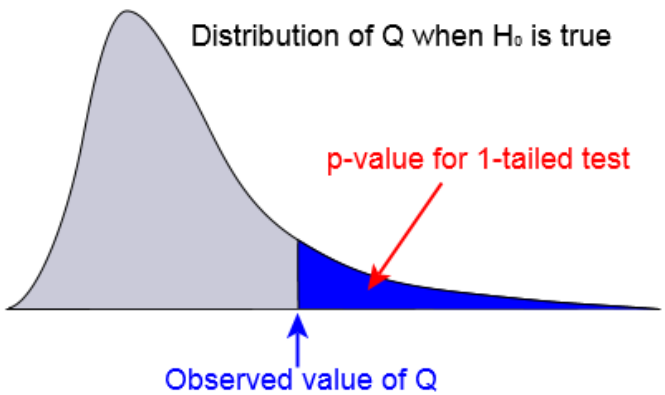
p-value	Interpretation
over 0.1	no evidence that H_0 does not hold
between 0.05 and 0.1	very weak evidence that H_0 does not hold
between 0.01 and 0.05	moderately strong evidence that H_0 does not hold
under 0.01	strong evidence that H_0 does not hold

P-value -summary

A hypothesis test is based on two competing hypotheses about the value of a parameter θ:

Null Hypothesis $H_0 : \theta = \theta_0$:

Alternative Hypothesis (one-tailed test) $H_A : \theta > \theta_0$



The hypothesis test is based on a test statistic that is some function of the data values: $Q = g(X_1, X_2, ..., X_n|\theta_0)$ whose distribution is fully known when H_0 is true (i.e. when θ_0 is the true parameter value). We evaluate the test statistic to assess whether it is `unusual` enough to throw doubt on the null hypothesis.

P-values close to zero throw doubt on the null hypothesis.

Fixed significance level (1)

Definition

The significance level is the probability of wrongly concluding that H_0 does not hold when it actually does.

One-tailed test

For example, it may be acceptable to have a 5% chance of concluding that $\theta < \theta_0$ when actually $\theta = \theta_0$. This means a significance level (tail area of the test statistic's distribution) of this test is $\alpha = 0.05$.

Two-tailed test

Values at both tails of the distribution of the test statistic result in rejection of H_0, so the corresponding tail areas should each have area $\alpha/2$ for a test with significance level α.

Fixed significance level (2)

Example

Cooking oil made from soybeans has little cholesterol and has been claimed to have only 15% saturated fat. A clinician believes that the saturated fat content is greater than 15% and randomly samples 13 bottles of soybean cooking oil for testing: 15.2 12.4 15.4 13.5 15.9 17.1 16.9 14.3 19.1 18.2 15.5 16.3 20.0.

$H_0 : \mu = 15$

$H_A : \mu > 15$

A significance level of $\alpha = 0.05$ means that the clinician is willing to wrongly conclude that the saturated fat content is over 15% when it really is 15% with probability 0.05.

The `t-statistic`: $t = \frac{\bar{x}-15}{s/\sqrt{13}} = 1.906$

The `rejection region`: $P(T > 1.782) = 0.05$

Conclusion: H_0 is rejected at the 5% significance level.

Question: What happens at 1% significance level when $P(T > 2.681) = 0.01$?

Type I and Type II errors

Type I error is the significance level of the test. The decision rule is usually defined to make the significance level 5% or 1%.
Type II error is wrongly accepting H_0 when it is false.

Instead of the probability of a Type II error, it is common to use the power of the test, defined as one minus the probability of a Type II error. The power of a test is the probability of correctly rejecting H_0 when it is false.

		Decision	
		accept H_0	reject H_0
Truth	H_0 is true		Significance level = P (Type I error)
	H_A (H_0 is false)	P (Type II error)	Power = 1 - P (Type II error)

Computer software can provide the p-value for a hypothesis test at 5% or 1% significance level (Type 1 error).

Power of a test (1)

It is clearly desirable to use a test whose power is as close to 1.0 as possible. There are three different ways to increase the power.

Increase the significance level

If the critical value for the test is adjusted, increasing the probability of a Type I error decreases the probability of a Type II error and therefore increases the power.

Use a different decision rule

For example, in a test about the mean of a normal population, a decision rule based on the sample median has lower power than a decision rule based on the sample mean.

Increase the sample size

By increasing the amount of data on which we base our decision about whether to accept or reject H_0, the probabilities of making errors can be reduced.

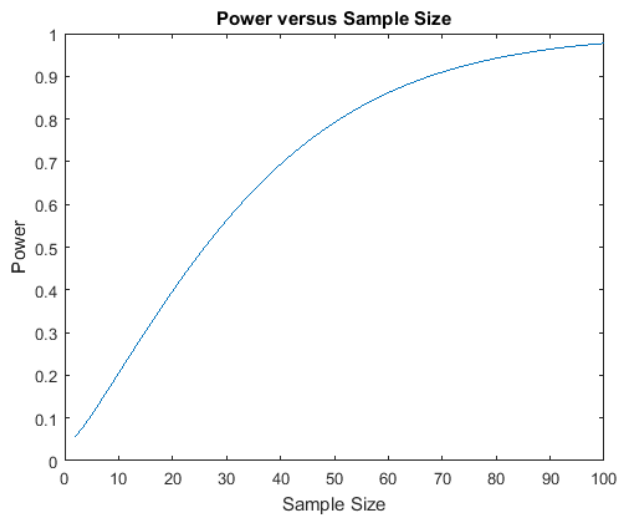
When the significance level is fixed, increasing the sample size is therefore usually the only way to improve the power.

Power of a test (2)

Ideally there should be the trade-off between low significance level (Type 1 error) and high power. The desired power of the test is usually 0.8.

The power of a test is not a single value since the alternative hypothesis allows for a range of different parameter values. It is represented by a power function that can be graphed against the possible parameter values.

MATLAB `sampsizepwr` can compute the sample size required to obtain a particular power for a hypothesis test, given the parameter value of the alternative hypothesis.



Shapiro-Wilk test and Kolmogorov-Smirnov tests use hypothesis testing

There are a number of statistical tests for assessing `normality`: Shapiro-Wilk test, Kolmogorov-Smirnov test, Jacque-Bera test etc.

The Shapiro-Wilk test can be used to verify whether data come from a normal distribution:
H_0 : sample data are not significantly different than a normal population.
H_A : sample data are significantly different than a normal population.
P-value > 0.05 mean the data are normal.
P-value < 0.05 mean the data are NOT normal.
Monte Carlo simulations proved the efficiency of Shapiro-Wilk test.
It is preferable that normality is assessed visually as well! The

Kolmogorov-Smirnov non-parametric test examines if scores are likely to follow some distribution in some population (not necessarily normal).

Grubbs' test for detecting outliers for normal distributions uses hypothesis testing

Test Statistic: The Grubbs' test statistic is defined as:

$$G = \frac{\max |Y_i - \bar{Y}|}{s}$$

with $\bar{Y}$ and s denoting the sample mean and standard deviation, respectively. The Grubbs' test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation.

This is the two-sided version of the test. The Grubbs' test can also be defined as one of the following one-sided tests:

1. test whether the minimum value is an outlier

$$G = \frac{\bar{Y} - Y_{min}}{s}$$

2. test whether the maximum value is an outlier

$$G = \frac{Y_{max} - \bar{Y}}{s}$$

Significance Level: α

Critical Region: For the two-sided test, the hypothesis of no outliers is rejected if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/(2N),N-2})^2}{N-2+(t_{\alpha/(2N),N-2})^2}}$$

with $t_{\alpha/(2N),N-2}$ denoting the critical value of the t distribution with $(N\text{-}2)$ degrees of freedom and a significance level of $\alpha/(2N)$.

For one-sided tests, we use a significance level of level of α/N.