Introductory Statistics

Tatiana Tyukina
tt51@leicester.ac.uk

Topic 3 - Hypothesis testing
3.1 - Inference based on two populations

- **Topic 0: Introduction**
- **Topic 1: Descriptive Statistics**
- **Topic 2: Estimators: Point estimation, Interval estimation**
  - 2.1 **Point estimation**,
  - 2.2 **Properties of estimators**,
  - 2.3 **Interval estimation**,
  - 2.4 **Inference based on other distributions**
- **Topic 3: Hypothesis Testing**
  - 3.1 Inference for two populations
  - 3.2 Hypothesis Testing Procedure
- Topic 4: Goodness of Fit: The χ^2 test
- Topic 5: Bayesian Estimation

Pivot RVs for population mean and proportion

$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$

$$T(\bar{X}, \mu) = \frac{\bar{X} - \mu}{\underline{\sigma/\sqrt{n}}}, \quad T(\bar{X}, \mu) \sim N(0, 1)$$

$$T\left(\frac{X}{n}, p\right) = \frac{X/n - p}{\sqrt{\underbrace{(X/n)(1 - (X/n))/n}_{\frac{1}{4n}}}}, \quad T\left(\frac{X}{n}, p\right) \sim N(0, 1)$$

OTHER PIVOT RVS

$$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

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

PIVOT RVS FOR PARAMETER OF EXPONENTIAL DISTRIBUTION

Let $X_1, X_2, ..., X_n$ be iid RVs, each with an exponential distribution with parameter β:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0$$

Then $Y = \sum_{i=1}^{n} X_i$ is a $Gamma(n, \beta)$ RV.

PIVOT RVS FOR PARAMETER OF EXPONENTIAL DISTRIBUTION

Let $X_1, X_2, ..., X_n$ be iid RVs, each with an exponential distribution with parameter β:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0$$

Then $Y = \sum_{i=1}^{n} X_i$ is a $Gamma(n, \beta)$ RV.

Consider the random variable $Q_i = \frac{2}{\beta} X_i$.
It can be shown, that the probability distribution of Q_i is $Gamma(1, 2)$
(Exercise: *to prove this consider cdf of* X, *change the variable under the integral and show that cdf of* Q *is* $\int_0^\infty 1/2 e^{-q/2} dq$).

Then the RV $T = \sum_{i=1}^{n} Q_i \sim Gamma(n, 2) = Gamma(2n/2, 2) = \chi^2_{2n}$.

$\chi^2_n = Gamma(\frac{n}{2}, 2)$

Pivot RVs for parameter of exponential distribution

Let $X_1, X_2, ..., X_n$ be iid RVs, each with an exponential distribution with parameter β:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0$$

Then $Y = \sum_{i=1}^n X_i$ is a $Gamma(n, \beta)$ RV.

Consider the random variable $Q_i = \frac{2}{\beta} X_i$.
It can be shown, that the probability distribution of Q_i is $Gamma(1,2)$ (Exercise: *to prove this consider cdf of X, change the variable under the integral and show that cdf of Q is $\int_0^\infty 1/2 e^{-q/2} dq$*).

Then the RV $T = \sum_{i=1}^n Q_i \sim Gamma(n,2) = Gamma(2n/2, 2) = \chi^2_{2n}$.

On the other hand, $T = \sum_{i=1}^n Q_i = \frac{2}{\beta} \cdot \sum_{i=1}^n X_i = \frac{2n}{\beta} \cdot \bar{X}$. $\sim \chi^2_{2n}$

PIVOT RVS FOR PARAMETER OF EXPONENTIAL DISTRIBUTION

Let $X_1, X_2, ..., X_n$ be iid RVs, each with an exponential distribution with parameter β (β is the "scale" parameter):

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0$$

$$T = \frac{2n}{\beta} \cdot \bar{X} \sim \chi^2_{2n}$$

Let $X_1, X_2, ..., X_n$ be iid RVs, each with an exponential distribution with parameter λ (λ is "rate" parameter):

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0$$

$$T = 2n\lambda \cdot \bar{X} \sim \chi^2_{2n}$$

MODULE STRUCTURE

- **Topic 0: Introduction**
- **Topic 1: Descriptive Statistics**
- **Topic 2: Estimators: Point estimation, Interval estimation**
  - 2.1 **Point estimation,**
  - 2.2 **Properties of estimators,**
  - 2.3 **Interval estimation,**
  - 2.4 **Inference based on other distributions**
- **Topic 3: Hypothesis Testing**
  - 3.1 Inference for two populations
  - 3.2 Hypothesis Testing Procedure
- Topic 4: Goodness of Fit: The χ^2 test
- Topic 5: Bayesian Estimation

Two random samples are selected from the two populations and the random variables (and observations) are independent from random variables within the same sample as well as from random variables from another sample.

In other words, the sample selected from one of the populations has no effect or bearing on the sample selected from the other population
and each possible pair of samples (one from one population and one from the other) is equally likely to be the pair of samples selected.

EXAMPLE

Two boxes of chocolate: milk and white

10 items of milk choc
10 items of white choc } only 3 m 3 w

with replacement
$10^3 \cdot 10^3$
1000 000

without replacement
$120 \cdot 120$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

Problem Statement:
Let $X_{1,1}, ..., X_{1,n_1}$ be a random sample from a distribution with mean μ_1 and variance σ_1^2 , and
let $X_{2,1}, ..., X_{2,n_2}$ be a random sample from a distribution with mean μ_2 and variance σ_2^2.

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

Problem Statement:
Let $X_{1,1}, ..., X_{1,n_1}$ be a random sample from a distribution with mean μ_1 and variance σ_1^2 , and
let $X_{2,1}, ..., X_{2,n_2}$ be a random sample from a distribution with mean μ_2 and variance σ_2^2.

Question: Obtain an estimate or a confidence interval for $\mu_1 - \mu_2$.

The first case: 2 samples from Normal distributions
Let $X_{1,1},...,X_{1,n_1}$ be a random sample from a normal distribution $N(\mu_1,\sigma_1^2)$, and let $X_{2,1},...,X_{2,n_2}$ be a random sample from a normal distribution $N(\mu_2,\sigma_2^2)$.

Let $\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}$ and $\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$.

As we assume that the two samples are independent, then $\bar{X}_1$ and $\bar{X}_2$ are independent, and the distribution of $\bar{X}_1 - \bar{X}_2$ is $N(\mu_1-\mu_2, \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2)$.
Now as in the one sample case, the confidence interval for $\mu_1-\mu_2$ is obtained as usual:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

The first case: 2 samples from Normal distributions
Let $X_{1,1},...,X_{1,n_1}$ be a random sample from a normal distribution $N(\mu_1,\sigma_1^2)$, and let $X_{2,1},...,X_{2,n_2}$ be a random sample from a normal distribution $N(\mu_2,\sigma_2^2)$.

Let $\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}$ and $\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$.

As we assume that the two samples are independent, then $\bar{X}_1$ and $\bar{X}_2$ are independent, and the distribution of $\bar{X}_1 - \bar{X}_2$ is $N(\mu_1-\mu_2, \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2)$.
Now as in the one sample case, the confidence interval for $\mu_1-\mu_2$ is obtained as usual:

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Interval Estimation for two population means

If we have two large samples:

- if σ_1^2 and σ_2^2 are **known**, then the large sample $(1-\alpha)100\%$ confidence interval for estimating $\mu_1 - \mu_2$:

$$\boxed{\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Interval Estimation for two population means

If we have two large samples:

- if σ_1^2 and σ_2^2 are **known**, then the large sample $(1-\alpha)100\%$ confidence interval for estimating $\mu_1 - \mu_2$:

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

- if σ_1^2 and σ_2^2 are **unknown**, then σ_1 and σ_2 can be replaced by respective sample standard deviations S_1 and S_2, then $n_1, n_2 \geq 30$;
thus the confidence interval is:

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{S_1^2/n_1 + S_2^2/n_2}$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

If we have small samples:

Main assumption:
The samples are from normal populations and are independent!

If we have small samples:

Main assumption:
The samples are from normal populations and are independent!

- if the two populations have unknown variance, but we can assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and the estimate of the variance can be obtained by pooling standard deviations of the two samples.

Definition

*The **pooled sample variance** S_p^2 is*

$$S_p^2 = \frac{\sum_{i=1}(X_{1,i} - \bar{X}_1)^2 + \sum_{i=1}(X_{2,i} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

Thus, when the two samples are independent,

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

Therefore, if the two sample are independent and are from two normal populations with equal variances the confidence interval for $\mu_1 - \mu_2$ is:

$$P(-t_{\alpha/2,n_1+n_2-2} < T < t_{\alpha/2,n_1+n_2-2}) = 1 - \alpha$$

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

Thus, when the two samples are independent,

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

Therefore, if the two sample are independent and are from two normal populations with equal variances the confidence interval for $\mu_1 - \mu_2$ is:

$$P(-t_{\alpha/2,n_1+n_2-2} < T < t_{\alpha/2,n_1+n_2-2}) = 1 - \alpha$$

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

EXAMPLE

Independent random samples from two normal populations with equal variances produced the following data.

Sample 1 : $1.2,\ 3.1,\ 1.7,\ 2.8,\ 3.0$

Sample 2 : $4.2,\ 2.7,\ 3.6,\ 3.9$

(a) Calculate the pooled estimate of σ^2.

(b) Obtain a 90% confidence interval for $\mu_1 - \mu_2$.

EXAMPLE

Independent random samples from two normal populations with equal variances produced the following data.
Sample 1 : 1.2, 3.1, 1.7, 2.8, 3.0
Sample 2 : 4.2, 2.7, 3.6, 3.9
(a) Calculate the pooled estimate of σ^2.
(b) Obtain a 90% confidence interval for $\mu_1 - \mu_2$.

Solution.

We have $n_1 = 5$, $n_2 = 4$. Also:

$$\bar{x}_1 = 2.36,\ s_1^2 = 0.733$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\bar{x}_2 = 3.6,\ s_2^2 = 0.420$$

Hence,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.599$$

Interval Estimation for two population means

Example

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.599$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

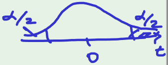
EXAMPLE

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.599$$

b) For the confidence coefficient 0.90, $\alpha = 0.10$ and from the t-table, $t_{0.05,7} = 1.895$.
Thus, a 90% confidence interval for $\mu_1 - \mu_2$ is

$n_1 = 5$
$n_2 = 4$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} =$$



$$(2.36 - 3.6) \pm 1.895 \cdot \sqrt{0.599 \left(\frac{1}{5} + \frac{1}{4} \right)} = -1.24 \pm 0.98$$

or $(-2.22, -0.26)$

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

In the case of small samples:

- if the equality of the variances cannot be reasonably assumed, $\sigma_1^2 \neq \sigma_2^2$, the previous procedure still can be used, except that the pivot random variable is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu,$$

where ν is the degree of freedom defined as (rounded down):

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

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

In the case of small samples:

- if the equality of the variances cannot be reasonably assumed, $\sigma_1^2 \neq \sigma_2^2$, the previous procedure still can be used, except that the pivot random variable is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu,$$

where ν is the degree of freedom defined as (rounded down):

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

$$\bar{X}_1 - \bar{X}_2 \pm \underline{t_{\alpha/2,\nu}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

INTERVAL ESTIMATION FOR TWO POPULATION MEANS

EXAMPLE

Assuming that the two populations are normally distributed with unknown and unequal variances. Two independent samples are taken with the following summary statistics:

$$n_1 = 16,\ \bar{x}_1 = 20.17,\ s_1 = 4.3$$

$$n_2 = 11,\ \bar{x}_2 = 19.23,\ s_2 = 3.8$$

Construct a 95% confidence interval for $\mu_1 - \mu_2$.

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$$

Answers:

$\nu = 23.312$; $\nu = 23$ df

the 95% confidence interval for $\mu_1 - \mu_2$:

$t_{0,025,23} = \pm 2.069$ $-2.3106 < \mu_1 - \mu_2 < 4.1906$

INTERVAL ESTIMATION FOR TWO POPULATION PROPORTIONS

In the large sample case:
Let X_1 and X_2 denote the numbers of successes observed in two independent sets of n_1 and n_2 Bernoulli trials, respectively, where p_1 and p_2 are the true success probabilities associated with each set of trials.
Recall that by the CLT

$$\frac{(X/n) - E(X/n)}{\sqrt{Var(X/n)/n}} \sim N(0,1)$$

It can be shown that

$$T = \frac{(X_1/n_1 - X_2/n_2) - (p_1 - p_2)}{\sqrt{\frac{X_1/n_1(1-X_1/n_1)}{n_1} + \frac{X_2/n_2(1-X_2/n_2)}{n_2}}} \sim N(0,1)$$

INTERVAL ESTIMATION FOR TWO POPULATION PROPORTIONS

In the large sample case:
Let X_1 and X_2 denote the numbers of successes observed in two independent sets of n_1 and n_2 Bernoulli trials, respectively, where p_1 and p_2 are the true success probabilities associated with each set of trials.
Recall that by the CLT

$$\frac{(X/n) - E(X/n)}{\sqrt{Var(X/n)/n}} \sim N(0,1)$$

It can be shown that

$$T = \frac{(X_1/n_1 - X_2/n_2) - (p_1 - p_2)}{\sqrt{\frac{X_1/n_1(1-X_1/n_1)}{n_1} + \frac{X_2/n_2(1-X_2/n_2)}{n_2}}} \sim N(0,1)$$

$$(X_1/n_1 - X_2/n_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{X_1/n_1(1 - X_1/n_1)}{n_1} + \frac{X_2/n_2(1 - X_2/n_2)}{n_2}}$$

Two samples are independent, large, n_1, n_2 and $(X_i/n_i)n_i > 5$ and $(1 - X_i/n_i)n_i > 5$ $i = 1, 2$.

INTERVAL ESTIMATION FOR TWO POPULATION PROPORTIONS

EXAMPLE

The phenomenon of handedness has been extensively studied in human populations. The percentages of adults who are right-handed, left-handed, and ambidextrous are well documented. What is not so well known is that a similar phenomenon is present in lower animals. Dogs, for example, can be either right-pawed or left-pawed. Suppose that in a random sample of 200 beagles, it is found that 55 are left-pawed and that in a random sample of 200 collies, 40 are left-pawed.
Obtain a 95% confidence interval for $p_1 - p_2$.

INTERVAL ESTIMATION FOR TWO POPULATION PROPORTIONS

EXAMPLE

The phenomenon of handedness has been extensively studied in human populations. The percentages of adults who are right-handed, left-handed, and ambidextrous are well documented. What is not so well known is that a similar phenomenon is present in lower animals. Dogs, for example, can be either right-pawed or left-pawed. Suppose that in a random sample of 200 beagles, it is found that 55 are left-pawed and that in a random sample of 200 collies, 40 are left-pawed.
Obtain a 95% confidence interval for $p_1 - p_2$.

Solution:

Let X_1 be a number of left-pawed beagles in the sample $n_1 = 200$, and
X_2 be a number of left-pawed collies in the sample $n_2 = 200$:

$$(X_1/n_1) = 55/200 = 0.275, \quad (X_2/n_2) = 40/200 = 0.2$$

$\frac{X_1}{n_1} \times n_1 > 5 \quad \left(1 - \frac{X_1}{n_1}\right) \cdot n_1 > 5$

continued...

. . .

Interval Estimation for two population proportions

Example

Solution (continued):

$$(X_1/n_1) = 55/200 = 0.275, \quad (X_2/n_2) = 40/200 = 0.2$$

The requirements are satisfied, hence, we can use approximation by standard normal distribution, so the 95% confidence interval for $p_1 - p_2$ is:

$$(X_1/n_1 - X_2/n_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{X_1/n_1(1 - X_1/n_1)}{n_1} + \frac{X_2/n_2(1 - X_2/n_2)}{n_2}} \Rightarrow$$

$$(0.275 - 0.2) \pm 1.96\sqrt{\frac{0.275 \cdot 0.725}{200} + \frac{0.2 \cdot 0.8}{200}} \Rightarrow$$

$$0.075 \pm 0.083$$

(−0,008 , 0,158)

Questions to take home

1. Define the notions of confidence interval;

Questions to take home

1. Define the notions of confidence interval;
2. Define the main properties of the pivot function;

QUESTIONS TO TAKE HOME

1. Define the notions of confidence interval;
2. Define the main properties of the pivot function;
3. Define pivot RV for two sample means if the samples are from normal distribution and variance is known;

QUESTIONS TO TAKE HOME

1. Define the notions of confidence interval;
2. Define the main properties of the pivot function;
3. Define pivot RV for two sample means if the samples are from normal distribution and variance is known;
4. Define pivot RV for two sample means if the samples are large and are from normal distribution and variance is unknown.