

STAT2201

Analysis of Engineering & Scientific Data

Unit 7

Slava Vaisman

The University of Queensland
School of Mathematics and Physics

Statistical inference (a reminder)

- ▶ Let $\mathbf{X}_1, \dots, \mathbf{X}_n \sim F(\mathbf{x})$ be a data drawn randomly from some **unknown** distribution F .
- ▶ Assume that the data is independent and identically distributed (i.i.d).
 1. $\mathbf{X}_i \sim F(\mathbf{x})$ for all $1 \leq i \leq n$
 2. \mathbf{X}_i s are independent
- ▶ Statistical Inference is the process of forming judgements about the parameters

Our setup

- ▶ Setup: A sample x_1, \dots, x_n (collected values).
- ▶ Model: An i.i.d. sequence of random variables, X_1, \dots, X_n .
- ▶ Parameter at question: The population mean, $\mathbb{E}[X_i]$.
- ▶ Point estimate: \bar{x} (described by the random variable \bar{X}).

The main objective: Devise hypothesis tests and confidence intervals for $\mu = \mathbb{E}[X_i]$.

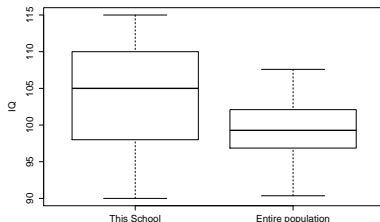
We distinguish between the two cases:

- ▶ Unrealistic (but simpler): The population variance, σ^2 , is known.
- ▶ More realistic: The variance is not known and estimated by the sample variance, s^2 .

Private school

Recall the private school example, which claims that its students have a higher IQ.

- ▶ Should we try to place our child in this school?
- ▶ Is the observed result *significant* (**can be trusted?**), or due to a *chance*?



The entire student population is known to have an IQ that is Gaussian distributed with mean 100 and variance 16.

Medical treatment

Recall experimental medical treatment example, in which 14 subjects were randomly assigned to control or treatment group. The survival times (in days) are shown in the table below.

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

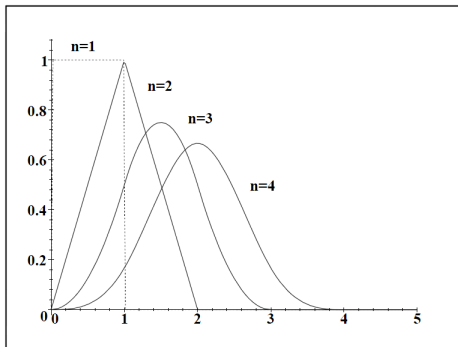
We asked:

- ▶ Did the treatment prolong the survival?
- ▶ Is the observed result *significant*, or due to a *chance*?

The variance is not known and estimated by the sample variance, s^2 .

Known variance — the Z-test

- ▶ A Z-test is any statistical test for which the distribution of the test statistic (the mean) under the null hypothesis can be approximated by a normal distribution (with known variance).
- ▶ Thanks to the central limit theorem, many test statistics are approximately **normally distributed** for large enough samples.



Z-test

- ▶ Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, (σ is known).
- ▶ Let us test $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.
- ▶ We choose the test statistics T to be $T = \bar{X}$.
- ▶ The p-value (the probability that under the null hypothesis, the random test statistic takes a value as extreme as or more extreme than the one observed) is

$$\text{p-value} = \mathbb{P}_{H_0} \left(\underbrace{\bar{X}}_{\text{random variable!}} > \underbrace{\bar{x}}_{\text{observed average!}} \right).$$

- ▶ Recall that: $p\text{-value low} \Rightarrow H_0 \text{ must go!}$

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	moderate evidence against H_0
$0.05 - 0.10$	suggestive evidence against H_0
> 0.1	little or no evidence against H_0

Z-test

- So, we need to calculate:

$$\text{p-value} = \mathbb{P}_{H_0} \left(\underbrace{\bar{X}}_{\text{random variable!}} > \underbrace{\bar{x}}_{\text{observed average!}} \right).$$

- Recall that If $X \sim N(\mu, \sigma^2)$, then

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

- Since \bar{X} is approximately normally distributed, we can standardize this normal random variable and arrive at the Z score:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}.$$

Z-test

We arrived at

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad (1)$$

since

$$\text{p-value} = \mathbb{P}_{H_0} \left(\underbrace{\bar{X}}_{\text{random variable!}} > \underbrace{\bar{x}}_{\text{observed average!}} \right)$$

$$= \mathbb{P}_{H_0} \left(\underbrace{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}_{(1)} < \underbrace{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}_{(1)} \right).$$

The Z-test

- ▶ Recall that (CLT)

$$\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\frac{\sigma}{n}}} \right) = \sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma} \right) \text{ approx. dist. } N(0, 1).$$

- ▶ For very small samples, the results we present are valid only if the population is normally distributed.
- ▶ We will generally require the sample size to be at least greater than 20.
- ▶ Let $H_0 : \mu = \mu_0$, and

$$H_1 : \begin{cases} \mu > \mu_0 & \text{right one sided test, or} \\ \mu < \mu_0 & \text{left one sided test, or} \\ \mu \neq \mu_0 & \text{two sided test} \end{cases}$$

- ▶ The test statistic is the average — \bar{X} .

The Z-test

So we define the Z-score, to be:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

► That is,

$$\mathbb{P}_{H_0} \left(\underbrace{\bar{X}}_{\text{test statistics}} > \underbrace{\bar{x}}_{\text{observed}} \right) = \mathbb{P}_{H_0} \left(\underbrace{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}_{Z \sim N(0,1)} > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right),$$

► or

$$\mathbb{P}_{H_0} (\bar{X} < \bar{x}) = \mathbb{P}_{H_0} \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right),$$

Types of tests

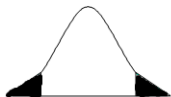
- ▶ **Right one-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \geq t)$.



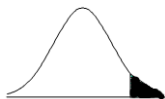
- ▶ **Left one-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \leq t)$.



- ▶ **Two-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \geq t) + \mathbb{P}_{H_0}(T \leq -t) = 2\mathbb{P}_{H_0}(T \geq t)$.



Right one-sided test ($H_1 : \mu \geq \mu_0 \rightarrow \mathbb{P}_{H_0}(T \geq t)$)



$$\mathbb{P}_{H_0}(\bar{X} > \bar{x}) = \mathbb{P}_{H_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \underbrace{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}_z\right) = 1 - \Phi(z)$$

Rejection Criterion for Fixed-Level Tests:

$$z > z_{1-\alpha}.$$

Left one-sided test ($H_1 : \mu \leq \mu_0 \rightarrow \mathbb{P}_{H_0}(T \leq t)$)



$$\mathbb{P}_{H_0}(\bar{X} < \bar{x}) = \mathbb{P}_{H_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \underbrace{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}_z\right) = \Phi(z)$$

Rejection Criterion for Fixed-Level Tests:

$$z < z_{\alpha}.$$

Two-sided test ($H_1 : \mu \neq \mu_0$ —
 $\mathbb{P}_{H_0}(T \geq |t|) + \mathbb{P}_{H_0}(T \leq -|t|)$)



$$\begin{aligned}\mathbb{P}_{H_0}(\bar{X} > |\bar{x}|) + \mathbb{P}_{H_0}(\bar{X} < -|\bar{x}|) &= 2\mathbb{P}_{H_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \underbrace{\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|}_z\right) \\ &= 2(1 - \Phi(|z|))\end{aligned}$$

Rejection Criterion for Fixed-Level Tests:

$$z < z_{\alpha/2} \quad \text{or} \quad z > z_{1-\alpha/2}.$$

Z-test summary

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

Model: $X_i \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ unknown but σ^2 known.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$

Alternative Hypotheses	P-value	Rejection Criterion for Fixed-Level Tests
$H_1 : \mu \neq \mu_0$	$P = 2[1 - \Phi(z)]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi(z)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi(z)$	$z < z_{\alpha}$

Z-test example (1)

```
using Distributions
using HypothesisTests
srand(12345)
```

```
private_school1 = rand(Normal(100,2), 50)
OneSampleZTest(private_school1,100)
```

```
private_school2 = rand(Normal(101,2), 50)
OneSampleZTest(private_school2,100)
```

Z-test example (2)

```
private_school1 = rand(Normal(100,2), 50)
OneSampleZTest(private_school1,100)
```

One sample z-test

Population details:

parameter of interest:	Mean
value under h_0 :	100
point estimate:	100.19550449696595
95% confidence interval:	(99.6332, 100.7577)

Test summary:

outcome with 95% confidence:	fail to reject h_0
two-sided p-value:	0.49553020954367355

Details:

number of observations:	50
z-statistic:	0.6815394561145689
population standard error:	0.28685719544473093

Z-test example (3)

```
private_school2 = rand(Normal(101,2), 50)
OneSampleZTest(private_school2,100)
```

One sample z-test

Population details:

parameter of interest:	Mean
value under h_0 :	100
point estimate:	100.80408350696453
95% confidence interval:	(100.26671, 101.34145)

Test summary:

outcome with 95% confidence:	reject h_0
two-sided p-value:	0.0033599975479617957

Details:

number of observations:	50
z-statistic:	2.9327264839267215
population standard error:	0.2741760990571197

Z-test's assumptions

- ▶ *Nuisance parameters* should be known, or estimated with high accuracy (standard deviation).
- ▶ In particular, when the sample size n is large you may use

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2},$$

instead of σ .

- ▶ The test statistic should follow a normal distribution. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

Z-test's assumptions

- ▶ In the (very realistic) case where σ^2 is not known, but rather estimated by S^2 , we would like to replace the test statistic, Z , with,

$$T = \frac{\bar{x} - \mu_0}{S/\sqrt{n}},$$

- ▶ Note that T no longer follows a Normal distribution!
- ▶ However, Under $H_0 : \mu = \mu_0$, and for moderate or large samples (e.g. $n > 100$) this statistic is approximately Normally distributed just like above. In this case, the procedures above work well.

But for smaller samples, the distribution of T is no longer Normally distributed. Nevertheless, it follows a well known and very famous distribution of classical statistics: The Student-t Distribution.

The t -test

- ▶ The t -statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland.



- ▶ It can happen that we do not know the standard deviation, or
- ▶ the number of samples is less than 30.

The t -test

In this case, use the t -test. The t statistics with $n - 1$ degrees of freedom is

$$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

where S is the estimated standard deviation:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- ▶ Use the t -test when the data is approximately normally distributed.
- ▶ For large n , t -test is indistinguishable from the z -test.

The t -distribution

- ▶ The probability density function of a Student-t Distribution with a parameter k , referred to as degrees of freedom, is,

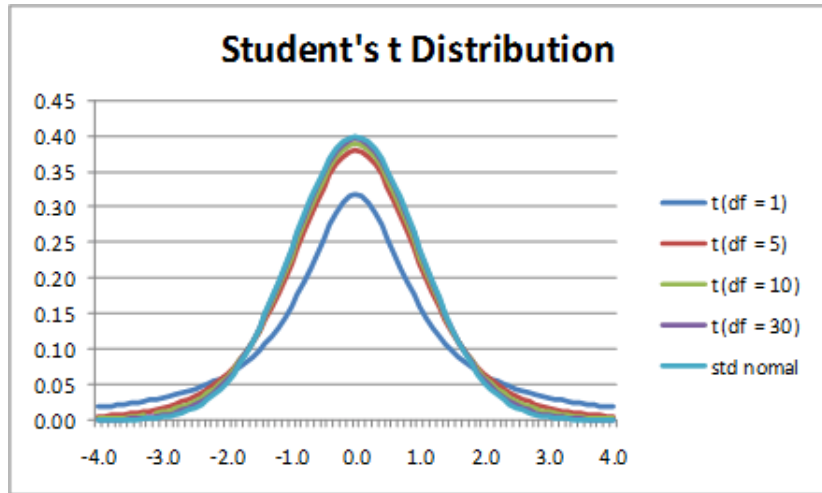
$$f(x, k) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k} \Gamma(k/2)} \frac{1}{[(x^2/k) + 1]^{(k+1)/2}}, \quad -\infty < x < \infty,$$

where $\Gamma(\cdot)$ is the Gamma-function:

$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx.$$

- ▶ It is a symmetric distribution about 0 and as $k \rightarrow \infty$, it approaches a standard Normal distribution.

The t -distribution



Why do we care about the t -distribution?

- ▶ Let X_1, X_2, \dots, X_n be an i.i.d. sample from a Normal distribution with mean μ and variance σ^2 .
- ▶ The random variable,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

has a t -distribution with $n - 1$ degrees of freedom.

- ▶ Now, knowing the distribution of T (and noticing it depends on the sample size, n), allows us to construct hypothesis tests and confidence intervals when σ^2 is not known, analogous to the (Z-tests and confidence intervals).

Confidence and prediction intervals

- ▶ If \bar{x} and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2 , a $100(1 - \alpha)$ confidence interval on μ is given by:

$$\bar{x} - t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}},$$

where $t_{1-\alpha/2, n-1}$ is the $1 - \alpha/2$ quantile of the t distribution with $n - 1$ degrees of freedom.

- ▶ A related concept is a $100(1 - \alpha)$ prediction interval (PI) on a single future observation from a normal distribution is given by

$$\bar{x} - t_{1-\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{1-\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}}$$

This is the range where we expect the $n + 1$ observation to be, after observing n observations and computing \bar{x} and s .

Types of tests (again)

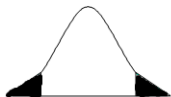
- ▶ **Right one-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \geq t)$.



- ▶ **Left one-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \leq t)$.



- ▶ **Two-sided test:** where H_0 is rejected for the p -value defined by $\mathbb{P}_{H_0}(T \geq t) + \mathbb{P}_{H_0}(T \leq -t) = 2\mathbb{P}_{H_0}(T \geq t)$.



t -test summary

Model: $X_i \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$

Alternative Hypotheses	P -value	Rejection Criterion for Fixed-Level Tests
$H_1 : \mu \neq \mu_0$	$P = 2[1 - F_{n-1}(t)]$	$t > t_{1-\alpha/2, n-1}$ or $t < t_{\alpha/2, n-1}$
$H_1 : \mu > \mu_0$	$P = 1 - F_{n-1}(t)$	$t > t_{1-\alpha, n-1}$
$H_1 : \mu < \mu_0$	$P = F_{n-1}(t)$	$t < t_{\alpha, n-1}$

t -test summary

- ▶ In the p-value calculation, $F_{n-1}(\cdot)$ denotes the CDF of the t -distribution with $n - 1$ degrees of freedom.
- ▶ As opposed to $\Phi(\cdot)$, the CDF of t is not tabulated in standard tables. So to calculate p-values, we use software.

t-test example (1)

```
private_school3 = [68.6869,88.7492,99.3467,81.4199 ]  
OneSampleZTest(private_school3,100)
```

One sample z-test

Population details:

parameter of interest:	Mean
value under h_0 :	100
point estimate:	84.550675
95% confidence interval:	(71.92434, 97.17700)

Test summary:

outcome with 95% confidence:	reject h_0
two-sided p-value:	0.01647705084278339

Details:

number of observations:	4
z-statistic:	-2.3981736722165747
population standard error:	6.442121010243833

t-test example (2)

```
private_school3 = [68.6869,88.7492,99.3467,81.4199 ]
```

```
OneSampleTTest(private_school3,100)
```

```
One sample t-test
```

```
-----
```

```
Population details:
```

```
    parameter of interest:    Mean
```

```
    value under h_0:         100
```

```
    point estimate:          84.550675
```

```
    95% confidence interval: (64.04897, 105.052379)
```

```
Test summary:
```

```
    outcome with 95% confidence: fail to reject h_0
```

```
    two-sided p-value:          0.09603209715776699
```

```
Details:
```

```
    number of observations:    4
```

```
    t-statistic:               -2.3981736722165747
```

```
    degrees of freedom:       3
```

```
    empirical standard error:  6.442121010243833
```