Comparison of two groups

The ASTA team

Contents

0.1 Response variable and explanatory variable

- We conduct an experiment, where we at random choose 50 IT-companies and 50 service companies and measure their profit ratio. Is there association between company type (IT/service) and profit ratio?
- In other words we compare samples from 2 different populations. For each company we register:
  - The binary variable `company type`, which is called **the explanatory variable** and divides data in 2 groups.
  - The quantitative variable `profit ratio`, which is called **the response variable**.

0.2 Dependent/independent samples

- In the example with profit ratio of 50 IT-companies and 50 service companies we have **independent samples**, since the same company cannot be in both groups.
- Now, think of another type of experiment, where we at random choose 50 IT-companies and measure their profit ratio in both 2009 and 2010. Then we may be interested in whether there is association between year and profit ratio?
- In this example we have **dependent samples**, since the same company is in both groups.
- Dependent samples may also be referred to as paired samples.

0.3 Comparison of two means (Independent samples)

- We consider the situation, where we have two quantitative samples:
  - Population 1 has mean μ_1, which is estimated by $\hat{\mu}_1 = \bar{y}_1$ based on a sample of size n_1.

  - Population 2 has mean μ_2, which is estimated by $\hat{\mu}_2 = \bar{y}_2$ based on a sample of size n_2.
  - We are interested in the difference $\mu_2 - \mu_1$, which is estimated by $d = \bar{y}_2 - \bar{y}_1$.
  - Assume that we can find the **estimated standard error** se_d of the difference and that this has degrees of freedom df.
  - Assume that the samples either are large or come from a normal population.
- Then we can construct a
  - confidence interval for the unknown population difference of means $\mu_2 - \mu_1$ by

    $$(\bar{y}_2 - \bar{y}_1) \pm t_{crit} se_d,$$

    where the critical t-score, t_{crit}, determines the confidence level.
  - significance test:
    - for the null hypothesis $H_0: \ \mu_2 - \mu_1 = 0$ and alternative hypothesis $H_a: \ \mu_2 - \mu_1 \neq 0$.
    - which uses the test statistic: $t_{obs} = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se_d}$, that has to be evaluated in a t-distribution with df degrees of freedom.

0.4 Comparison of two means (Independent samples)

- In the independent samples situation it can be shown that

  $$se_d = \sqrt{se_1^2 + se_2^2},$$

  where se_1 and se_2 are estimated standard errors for the sample means in populations 1 and 2, respectively.
- We recall, that for these we have $se = \frac{s}{\sqrt{n}}$, i.e.

  $$se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

  where s_1 and s_2 are estimated standard deviations for population 1 and 2, respectively.
- **The degrees of freedom** df for se_d can be estimated by a complicated formula, which we will not present here.
- For the confidence interval and the significance test we note that:
  - If both n_1 and n_2 are above 30, then we can use the standard normal distribution (z-score) rather than the t-distribution (t-score).
  - If n_1 or n_2 are below 30, then we let **Python** calculate the degrees of freedom and p-value/confidence interval.

0.5 Example: Comparing two means (independent samples)

We return to the `Chile` data. We study the association between the variables `sex` and `statusquo` (scale of support for the status-quo). So, we will perform a significance test to test for difference in the mean of `statusquo` for male and females.

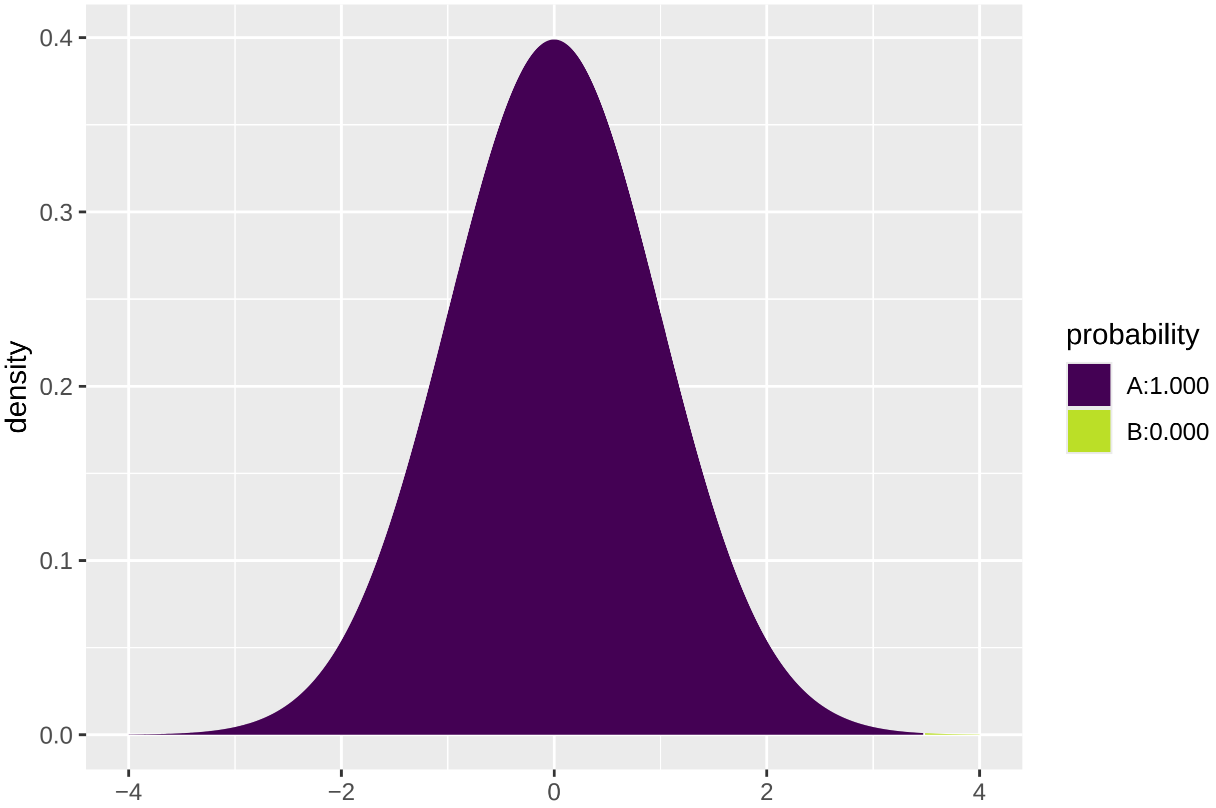
```
import pandas as pd

Chile = pd.read_csv("https://asta.math.aau.dk/datasets?file=Chile.txt", sep = "\t")

stats = Chile.groupby("sex")["statusquo"].agg(
    ['mean',
     'std',
     'count'
]).reset_index() # reset_index() resets the grouping again
stats
```

```
##   sex      mean       std  count
## 0   F  0.065706  1.003212   1368
## 1   M -0.068355  0.992803   1315
```

- Difference: $d = 0.0657 - (-0.0684) = 0.1341$.
- Estimated standard deviations: $s_1 = 1.0032$ (females) and $s_2 = 0.9928$ (males).
- Sample sizes: $n_1 = 1368$ and $n_2 = 1315$.
- Estimated standard error of difference: $se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.0032^2}{1368} + \frac{0.9928^2}{1315}} = 0.0385$.
- Observed t-score for $H_0: \ \mu_1 - \mu_2 = 0$ is: $t_{obs} = \frac{d-0}{se_d} = \frac{0.1341}{0.0385} = 3.4786$.
- Since both sample sizes are "pretty large" (> 30), we can use the z-score instead of the t-score for finding the p-value (i.e. we use the standard normal distribution):



```
from scipy.stats import norm
1 - norm.cdf(x = 3.4786, loc = 0, scale = 1)
```

```
## np.float64(0.00025202016841718855)
```

- Then the p-value is $2 \cdot 0.00025 = 0.0005$, so we reject the null hypothesis.
- We can leave all the calculations to the software:

```
from scipy.stats import ttest_ind

female = Chile.loc[Chile['sex'] == "F", 'statusquo'].dropna()
male = Chile.loc[Chile['sex'] == "M", 'statusquo'].dropna()

stat, pval = ttest_ind(female, male, equal_var = False)

print("t-statistic:", stat)
```

```
## t-statistic: 3.4785834945762018
```

```
print("p-value:", pval)
```

```
## p-value: 0.0005121107038059029
```

- We recognize the t-score 3.4786 and the p-value 0.0005. The estimated degrees of freedom $df = 2679$ is so large that we can not tell the difference between results obtained using z-score and t-score.

0.6 Comparison of two means: confidence interval (independent samples)

- We have already found all the ingredients to construct a **confidence interval for** $\mu_2 - \mu_1$:
  - $d = \bar{y}_2 - \bar{y}_1$ estimates $\mu_2 - \mu_1$.
  - $se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ estimates the standard error of d.
- Then:

$$d \pm t_{crit} se_d$$

  is a confidence interval for $\mu_2 - \mu_1$.
- The critical t-score, t_{crit} is chosen corresponding to the wanted confidence level. If n_1 and n_2 both are greater than 30, then $t_{crit} = 2$ yields a confidence level of approximately 95%.

0.7 Comparison of two means: paired t-test (dependent samples)

- Experiment:
  - You choose 32 students at random and measure their average reaction time in a driving simulator while they are listening to radio or audio books.
  - Later the same 32 students redo the simulated driving while talking on a cell phone.
- It is interesting to investigate whether or not the fact that you are actively participating in a conversation changes your average reaction time compared to when you are passively listening.
- So we have 2 samples corresponding to with/without phone. In this case we have **dependent** samples, since we have 2 measurement for each student.
- We use the following strategy for analysis:
  - For each student calculate **the change** in average reaction time with and without talking on the phone.
  - The changes $d_1, d_2, \ldots, d_{32}$ are now considered as **ONE** sample from a population with mean μ.
  - Test the hypothesis $H_0 : \mu = 0$ as usual (using a t-test for testing the mean as in the previous lecture).

0.7.1 Reaction time example

- Data is organized in a data frame with 3 variables:
  - `student` (integer – a simple id)
  - `reaction_time` (numeric – average reaction time in milliseconds)
  - `phone` (factor – `yes`/`no` indicating whether speaking on the phone)

```
reaction = pd.read_csv("https://asta.math.aau.dk/datasets?file=reaction.txt", sep = "\t")
reaction.head(3)
```

```
##    student  reaction_time phone
## 0        1            604    no
## 1        2            556    no
## 2        3            540    no
```

Instead of doing manual calculations we let the software perform the significance test (using a paired test as our samples are paired/dependent):

```
from scipy.stats import ttest_rel

yes = reaction[reaction['phone'] == "yes"]
no  = reaction[reaction['phone'] == "no"]
print(all(yes['student'].values == no['student'].values))
```

```
## True
```

```
stat, pval = ttest_rel(no["reaction_time"], yes["reaction_time"])
print("t-statistic:", stat)
```

```
## t-statistic: -5.456300665835772
```

```
print("p-value:", pval)
```

```
## p-value: 5.803405318112956e-06
```

- With a p-value of 0.0000058 we reject that speaking on the phone has no influence on the reaction time.
- To understand what is going on, we can manually find the reaction time difference for each student and do a one sample t-test on this difference:

```
from scipy.stats import ttest_1samp
diff = no["reaction_time"].values - yes["reaction_time"].values
stat, pval = ttest_1samp(diff, popmean = 0)
print("t-statistic:", stat)
```

```
## t-statistic: -5.456300665835772
```

```
print("p-value:", pval)
```

```
## p-value: 5.803405318112956e-06
```

1 Comparison of two proportions

1.1 Comparison of two proportions

- We consider the situation, where we have two qualitative samples and we investigate whether a given property is present or not:
  - Let the proportion of population 1 which has the property be π_1, which is estimated by $\hat{\pi}_1$ based on a sample of size n_1.
  - Let the proportion of population 2 which has the property be π_2, which is estimated by $\hat{\pi}_2$ based on a sample of size n_2.
  - We are interested in the difference $\pi_2 - \pi_1$, which is estimated by $d = \hat{\pi}_2 - \hat{\pi}_1$.
  - Assume that we can find the **estimated standard error** se_d of the difference.
- Then we can construct
  - an approximate confidence interval for the difference, $\pi_2 - \pi_1$.
  - a significance test.

1.2 Comparison of two proportions: Independent samples

- In the situation where we have independent samples we know that

$$se_d = \sqrt{se_1^2 + se_2^2},$$

  where se_1 and se_2 are the estimated standard errors for the sample proportion in population 1 and 2, respectively.
- We recall, that these are given by $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, i.e.

$$se_d = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

- A (approximate) confidence interval for $\pi_2 - \pi_1$ is obtained by the usual construction:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z_{crit} se_d,$$

  where the critical z-score determines the confidence level.

1.3 Approximate test for comparing two proportions (independent samples)

- We consider the null hypothesis H_0: $\pi_1 = \pi_2$ (equivalently $H_0 : \pi_1 - \pi_2 = 0$) and the alternative hypothesis H_a: $\pi_1 \neq \pi_2$.
- Assuming H_0 is true, we have a common proportion π, which is estimated by

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2},$$

  i.e. we aggregate the populations and calculate the relative frequency of the property (with other words: we estimate the proportion, π, as if the two samples were one).
- Rather than using the estimated standard error of the difference from previous, we use the following that holds under H_0:

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- The observed test statistic/z-score for H_0 is then:

$$z_{obs} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{se_0},$$

  which is evaluated in the standard normal distribution.
- The p-value is calculated in the usual way.

WARNING: The approximation is only good, when $n_1\hat{\pi},\ n_1(1-\hat{\pi}),\ n_2\hat{\pi},\ n_2(1-\hat{\pi})$ all are greater than 5.

1.4 Example: Approximate confidence interval and test for comparing proportions

We return to the `Chile` dataset. We make a new binary variable indicating whether the person intends to vote no or something else (and we remember to tell the software that it should think of this as a grouping variable):

```
import numpy as np

Chile["vote"].value_counts()
```

```
## vote
## N    889
## Y    868
## U    588
## A    187
## Name: count, dtype: int64
```

```
Chile["vote"].value_counts(dropna = False)
```

```
## vote
## N      889
## Y      868
## U      588
## A      187
## NaN    168
## Name: count, dtype: int64
```

```
# Step 1: initialize with NA
voteNo = pd.Series([np.nan] * len(Chile), dtype = "object")
# Step 2: mark TRUE where vote == "N"
voteNo[Chile["vote"] == "N"] = True
# Step 3: mark FALSE where vote != "N" (but not NA)
voteNo[(Chile["vote"].notna()) & (Chile["vote"] != "N")] = False

Chile["voteNo"] = pd.Categorical(
    voteNo,
    categories=[True, False],
    ordered=False
)
Chile["voteNo"].value_counts(dropna = False)
```

```
## voteNo
## False    1643
## True      889
## NaN       168
## Name: count, dtype: int64
```

We study the association between the variables `sex` and `voteNo`:

```
tab = pd.crosstab(Chile["sex"], Chile["voteNo"], dropna = True)
tab
```

```
## voteNo  True  False
## sex
## F        363    946
## M        526    697
```

This gives us all the ingredients needed in the hypothesis test:

- Estimated proportion of men that vote no: $\hat{\pi}_1 = \frac{526}{526+697} = 0.430$
- Estimated proportion of women that vote no: $\hat{\pi}_2 = \frac{363}{363+946} = 0.277$

1.5 Example: Approximate confidence interval (cont.)

- Estimated difference:

$$d = \hat{\pi}_2 - \hat{\pi}_1 = 0.277 - 0.430 = -0.153$$

- Standard error of difference:

$$\begin{aligned} se_d &= \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \\ &= \sqrt{\frac{0.430(1-0.430)}{1223} + \frac{0.277(1-0.277)}{1309}} = 0.0188. \end{aligned}$$

- Approximate 95% confidence interval for difference:

$$d \pm 1.96 \cdot se_d = (-0.190, -0.116).$$

1.6 Example: p-value (cont.)

- Estimated common proportion:

$$\hat{\pi} = \frac{1223 \times 0.430 + 1309 \times 0.277}{1309 + 1223} = \frac{526 + 363}{1309 + 1223} = 0.351.$$

- Standard error of difference when $H_0: \ \pi_1 = \pi_2$ is true:

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 0.0190.$$

- The observed test statistic/z-score:

$$z_{obs} = \frac{d}{se_0} = -8.06.$$

- The test for H_0 against $H_a: \pi_1 \neq \pi_2$ yields a p-value that is practically zero, i.e. we can reject that the proportions are equal.

1.7 Automatic calculation

```
from statsmodels.stats.proportion import proportions_ztest
from statsmodels.stats.proportion import confint_proportions_2indep

Chile2 = Chile.dropna(subset=["voteNo"])

counts = Chile2.groupby("sex")["voteNo"].apply(lambda x: (x == True).sum())
nobs   = Chile2.groupby("sex")["voteNo"].count()

print("Counts (successes):\n", counts)
```

```
## Counts (successes):
##  sex
## F    363
## M    526
## Name: voteNo, dtype: int64
```

```
print("Totals (nobs):\n", nobs)
```

```
## Totals (nobs):
##  sex
## F    1309
## M    1223
## Name: voteNo, dtype: int64
```

```
print("sample estimates:\n")
```

```
## sample estimates:
```

```
print(counts/nobs)
```

```
## sex
## F    0.277311
## M    0.430090
## Name: voteNo, dtype: float64
```

```
# Two-sample proportion z-test
stat, pval = proportions_ztest(count=counts, nobs=nobs, value=0, alternative='two-sided')

ci_low, ci_high = confint_proportions_2indep(
    count1=counts.iloc[0], nobs1=nobs.iloc[0],
    count2=counts.iloc[1], nobs2=nobs.iloc[1],
    method="wald", alpha=0.05
)
```

```
print("p-value:", pval)
```

```
## p-value: 8.389098566796607e-16
```

```
print(f"95% CI for difference: ({ci_low:.3f}, {ci_high:.3f})")
```

```
## 95% CI for difference: (-0.190, -0.116)
```

1.8 Fisher's exact test

- If $n_1\hat{\pi}$, $n_1(1-\hat{\pi})$, $n_2\hat{\pi}$, $n_2(1-\hat{\pi})$ are not all greater than 5, then the approximate test cannot be trusted. Instead you can use Fisher's exact test:

```
from scipy.stats import fisher_exact

oddsratio, pvalue = fisher_exact(tab)
print("Odds ratio:", oddsratio)
```

```
## Odds ratio: 0.5084667079317358
```

```
print("p-value:", pvalue)
```

```
## p-value: 1.0396837491279301e-15
```

- Again the p-value is seen to be extremely small, so we definitely reject the null hypothesis of equal `voteNo` proportions for women and men.

1.9 Agresti: Overview of comparison of two groups

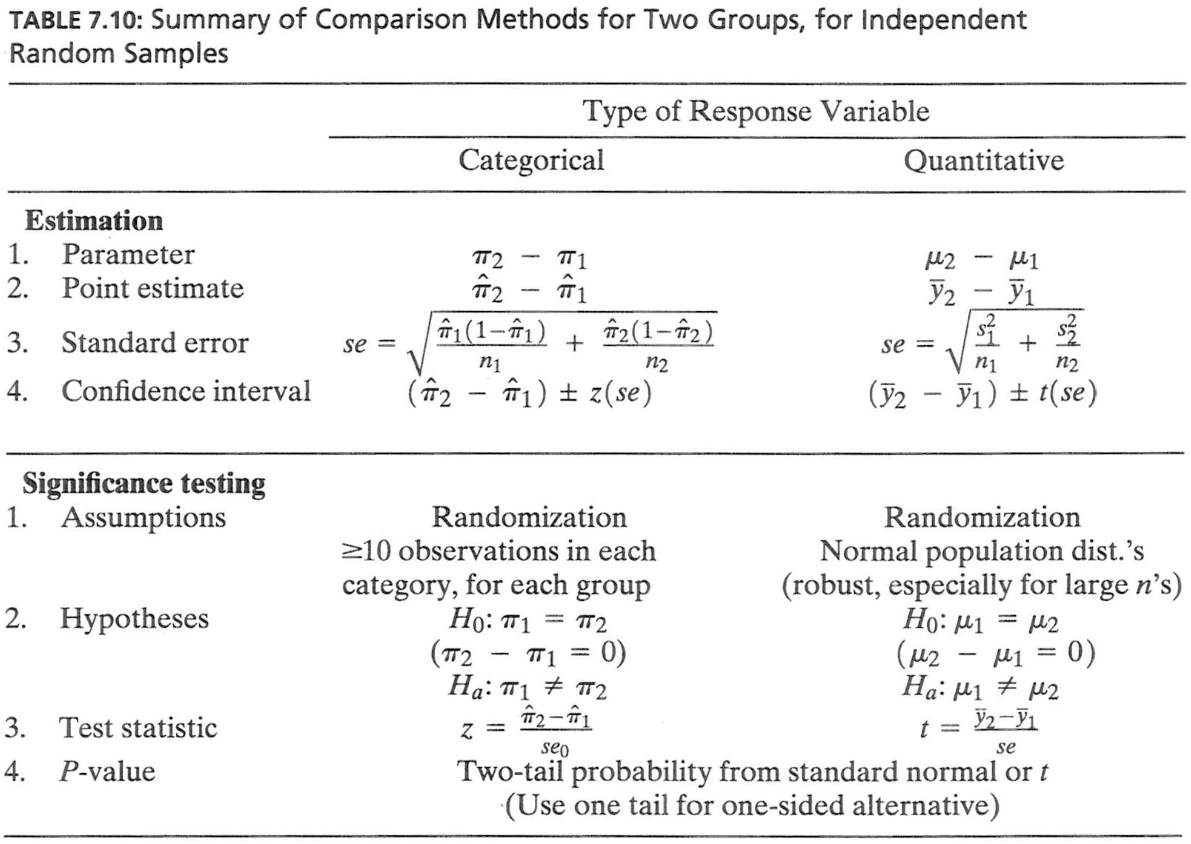


TABLE 7.10: Summary of Comparison Methods for Two Groups, for Independent Random Samples

	Type of Response Variable	
	Categorical	Quantitative
Estimation		
1. Parameter	$\pi_2 - \pi_1$	$\mu_2 - \mu_1$
2. Point estimate	$\hat{\pi}_2 - \hat{\pi}_1$	$\bar{y}_2 - \bar{y}_1$
3. Standard error	$se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$	$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Confidence interval	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$	$(\bar{y}_2 - \bar{y}_1) \pm t(se)$
Significance testing		
1. Assumptions	Randomization ≥10 observations in each category, for each group	Randomization Normal population dist.'s (robust, especially for large n's)
2. Hypotheses	H_0: $\pi_1 = \pi_2$ $(\pi_2 - \pi_1 = 0)$ H_a: $\pi_1 \neq \pi_2$	H_0: $\mu_1 = \mu_2$ $(\mu_2 - \mu_1 = 0)$ H_a: $\mu_1 \neq \mu_2$
3. Test statistic	$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0}$	$t = \frac{\bar{y}_2 - \bar{y}_1}{se}$
4. *P*-value	Two-tail probability from standard normal or t (Use one tail for one-sided alternative)	