Chapter 7: Inferences for Distributions

Halima Bensmail

CS502

Monday 4-7pm

LAS Hall C

- The previous chapter introduces confidence intervals and hypothesis tests for the mean – assuming σ is known (or that the difference between s and σ is small enough to be ignored)

- This was done to simplify the discussion and the calculations.

- The procedures we will use most often are covered in Chapter 7 – where we do not assume that σ is known.

Z- or t-test?

In the previous chapter we studied procedures based on the Z- or normal distribution. These techniques assume:

•The distribution of the mean is normal*

•The standard deviation of the population is known

In most applications the SD is not known but estimated from a sample. In this case the inferential procedures need to be adjusted. In this chapter we will study the t-distribution and some of its uses in inferential procedures.

*This can be achieved if (1) the population has a normal distribution or, (2) the sample size is large.

The Standard Error of the Sample Mean

STANDARD ERROR

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

A very useful distribution, particularly for small n

THE t DISTRIBUTIONS

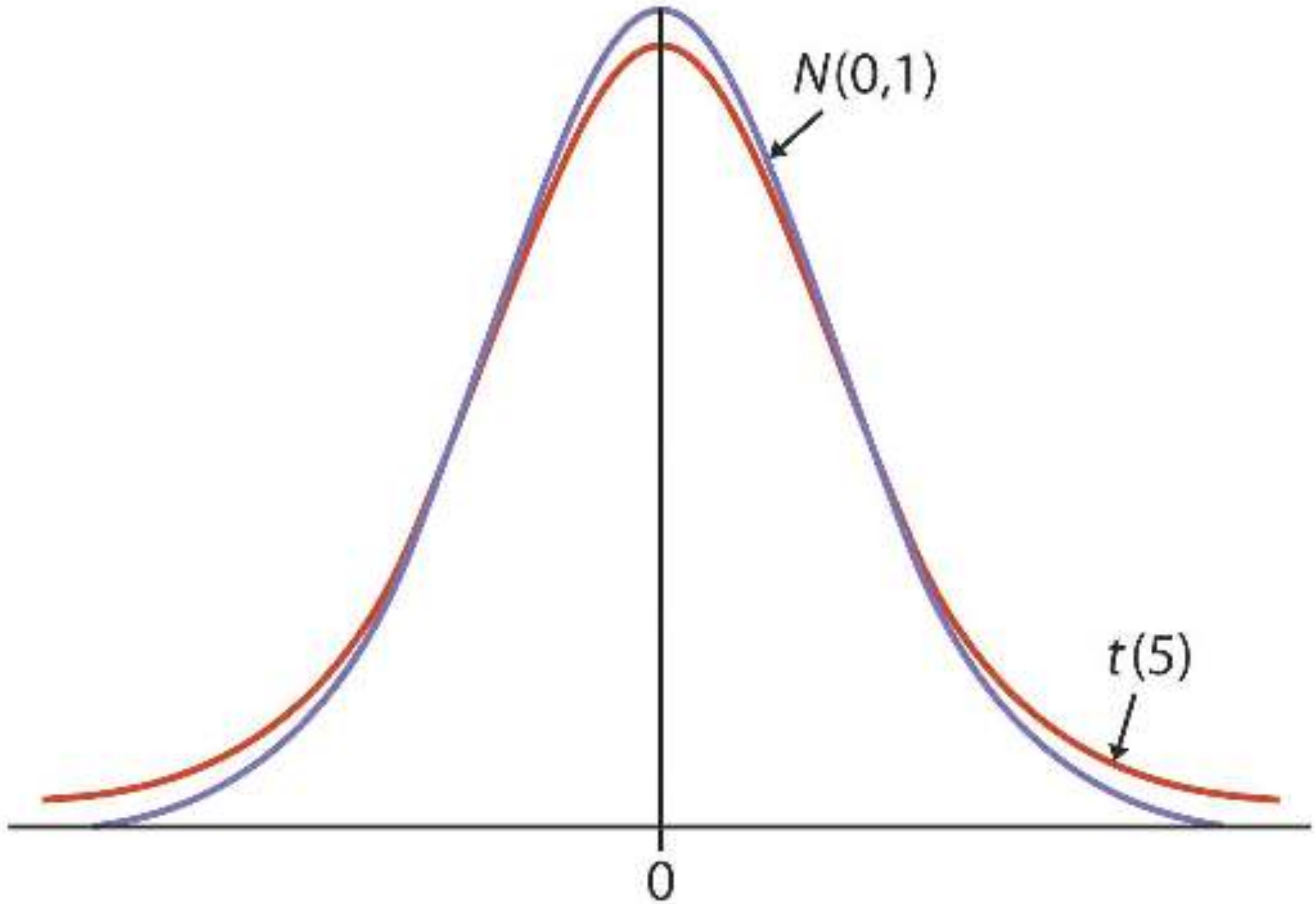
Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample t statistic**

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

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

The **t** distribution resembles the normal distribution, but is slightly more variable than Normal(0,1) due to the variation in the denominator of **t**.

For relatively large n, the t-distribution is very close to the standard normal



When the standard deviation is **unknown**, inferences about the mean should be based on the **t-distribution**.

THE ONE-SAMPLE t CONFIDENCE INTERVAL

Suppose that an SRS of size n is drawn from a population having unknown mean μ. A level C **confidence interval** for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The margin of error is

$$t^* \frac{s}{\sqrt{n}}$$

This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

Remember this!

THE ONE-SAMPLE t TEST

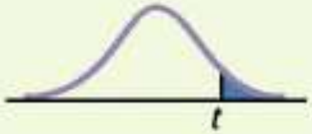
Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

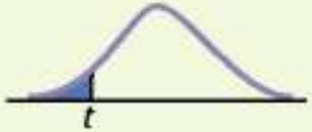
Remember this too!

In terms of a random variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

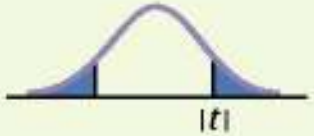
$H_a: \mu > \mu_0$ is $P(T \geq t)$



$H_a: \mu < \mu_0$ is $P(T \leq t)$



$H_a: \mu \neq \mu_0$ is $2P(T \geq |t|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

More realistic situation: σ is unknown.

Approach: Replace by estimate $\hat{\sigma} = s$

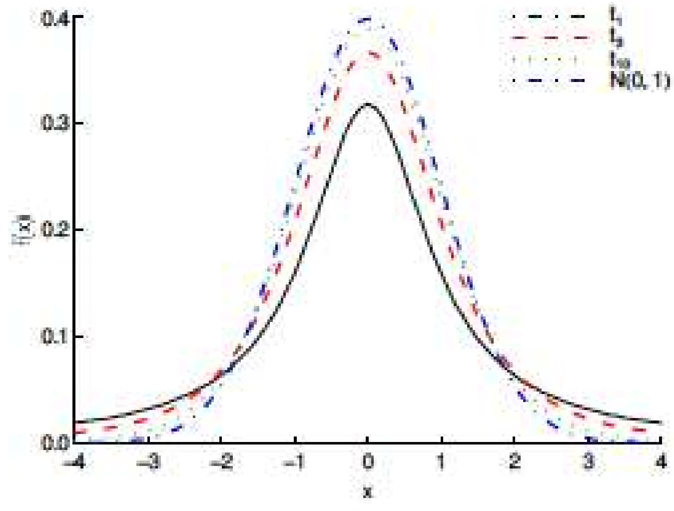
This approach leads to the $\boldsymbol{t}$ **statistic**

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

It is t distributed with $n - 1$ degrees of freedom.

Notation: Critical values of distributions

z_α	standard normal distribution
$t_{n,\alpha}$	t distribution with n degrees of freedom



Confidence interval for the mean μ (σ unknown)

The interval

$$\left[\bar{X} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$$

is a $(1 - \alpha)$ confidence interval for the mean μ.

Example: Cholesterol levels

In the study on cholesterol levels, the standard deviation of the decrease of cholesterol level was unknown.

- $\hat{\mu}_D = 36.89$, $\hat{\sigma}_D = 50.94$
- $t_{27,0.025} = 2.05$
- Then

$$\left[36.89 - 2.05 \cdot \frac{50.94}{\sqrt{27}}, 36.89 + 2.05 \cdot \frac{50.94}{\sqrt{27}}\right] = [16.78, 57.01]$$

  is a 95% confidence interval for μ_D
- The large sample confidence interval based on (*) was [18.00,55.78].

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Example: Level of vitamin C

The following data are the amounts of vitamin C, measured in milligrams per 100 grams (mg/100 g) of corn soy blend, for a random sample of size 8 from a production run:

$$26 \quad 31 \quad 23 \quad 22 \quad 11 \quad 22 \quad 14 \quad 31$$

What is the 95% confidence interval for μ, the mean vitamin C content of the CSB produced during this run?

- $\hat{\mu} = 22.5$, $\hat{\sigma} = 7.2$, $t_{7,0.025} = 2.36$
- The 95% confidence interval for μ is

$$\left[22.5 - \frac{2.36 \cdot 7.2}{\sqrt{8}}, 22.5 + \frac{2.36 \cdot 7.2}{\sqrt{8}}\right] = [16.5, 28.5].$$

- The large sample CI would be $[17.5, 27.5]$.

Example

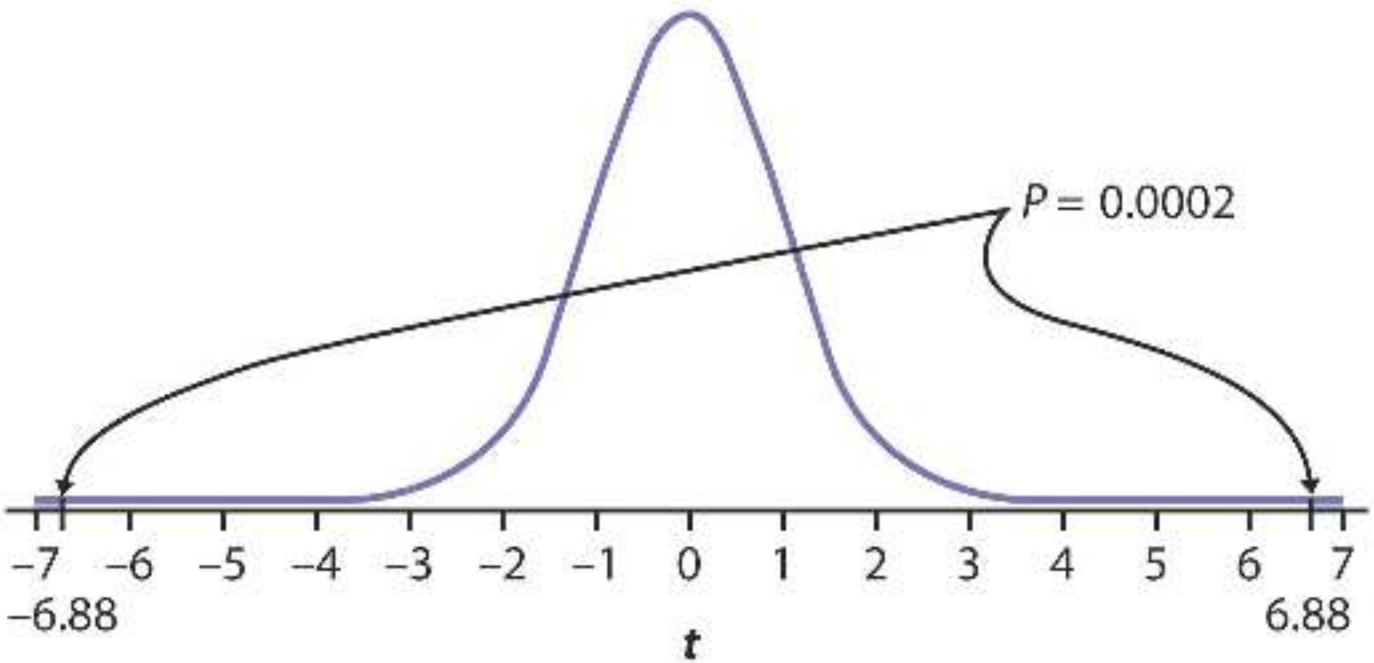
- Suppose the null hypothesis is we want to test the mean vitamin C of the Corn Soy Blend is:
  - H_0: μ = 40 versus H_A μ # 40
- What are the data? n = 8, mean = 22.5, s = 7.2
- Calculate the test statistic t?
- Test statistic: t = -6.88 Interpret:
- Now compute the p-value (using t-table with
- How many degrees of freedom?
- Make a conclusion

Two-sided p-value

Use t table with n-1 = 7, since this reflects the information in s.

Using any software, we obtain this probability exactly:

$P(t < -6.88) = P(t > 6.88) = .0001$

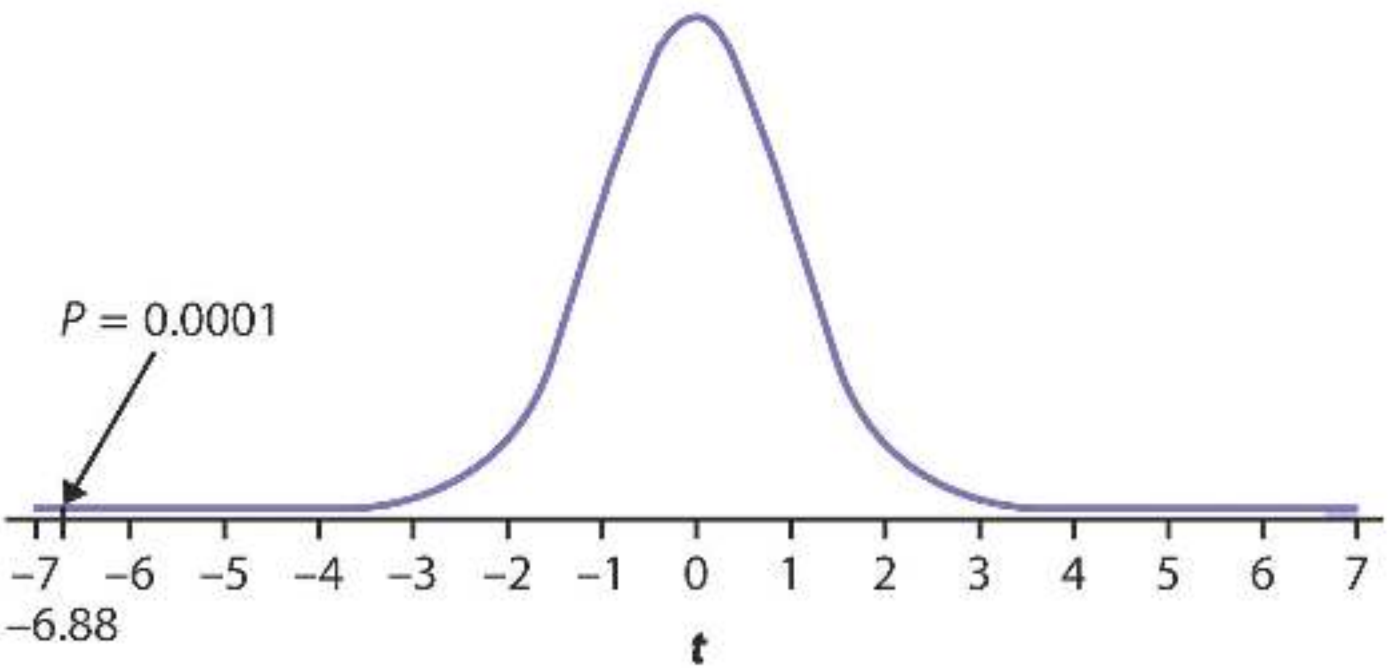


For this two-sided test, p-value = $2P(t > 6.88) = .0002$

Interpretation: We reject H_0 and conclude that the vitamin C Content for this run is below the specifications

If, in advance of collecting data, we were only concerned about
The possibility of μ being smaller than 40, we would have H_a: $\mu < 40$.
The p-value is the probability of getting an outcome that is even more supportive of the alternative hypothesis.

One-sided p-value



<=Supportive of H_a Not supportive of H_a=>

Student test or T test in R

Description

Density, distribution function, quantile function and random generation for the t distribution with df degrees of freedom (and optional non-centrality parameter ncp).

Usage

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

The Accuracy of Radon Detectors

Radon is a radioactive gas that is released from the decay of uranium.

How accurate are radon detectors of a type sold to home owners? To answer this question, university researchers placed **12** detectors in a chamber that exposed them to **105** picocuries per liter (pCi/l) of radon. The readings were as follows:

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

Is there evidence to suggest that the mean reading of all detectors of this type **differs** from the true value of 105?

The Accuracy of Radon Detectors

Test Mean=value

Hypothesized Value	105
Actual Estimate	104.133
df	11
Std Dev	9.39742

	t Test
Test Statistic	-0.3195
Prob > |t|	0.7554
Prob > t	0.6223
Prob < t	0.3777

Which of these p-values do we use?

95% Confidence Interval for μ: the mean reading (@ 105pCi/l)

- $104.133 +/- (t_{.025, n-1})\ 9.4/\sqrt{12}$
- Find the correct t value in t table or R
- Interpret this interval

Since statistical procedures depend on assumptions, it is better if they are not very sensitive to departures from those assumptions

ROBUST PROCEDURES

A statistical inference procedure is called **robust** if the probability calculations required are insensitive to violations of the conditions that usually justify the procedure.

T-tests are rather robust, as long as the degrees of freedom are not too small.

Two samples Inferences

Important Topic

- Two-Sample Comparison of Means
- The two samples will either be:

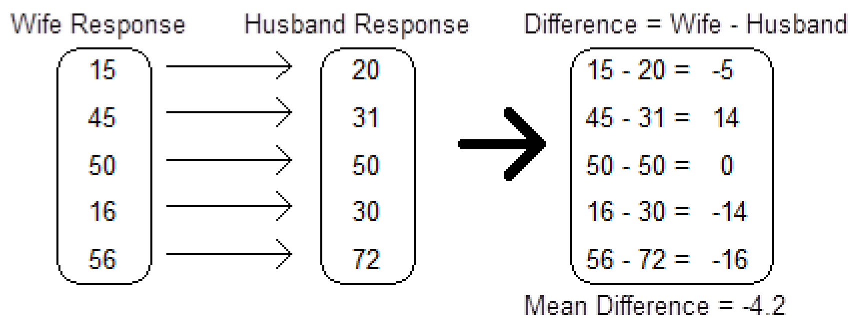
1. Matched in Pairs
2. Independent

Matched pairs is both more effective and simpler to analyze than independent samples. We begin there.

Definition

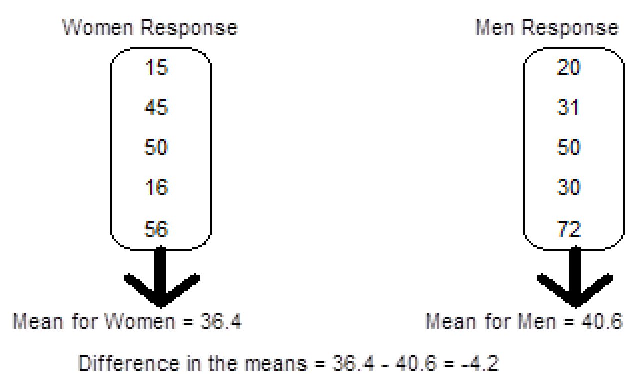
- Matched Pairs

With paired data, we are interested in comparing the responses within each pair. We will analyze the differences of the responses that form each pair.



- Independent Pairs

In the two independent samples scenario, we will compare the responses of one treatment group as a whole to the responses of the other treatment group as a whole. We will calculate summary measures for the observations from one treatment group and compare them to similar summary measures calculated from the observations from the other treatment group.



The Effect of Language Instruction

A company contracts with a language institute to provide individualized instruction in foreign languages for its executives who will be posted overseas. Is this instruction effective?

Last year, **20** executives studied French. All had some knowledge of French, so they were given the **Modern Language Association's listening test** of understanding of spoken French before their instruction began. After several weeks, the executives took the listening **test again**.

Analyze this data as a matched pairs situation.

Matched Pairs

TABLE 7.2 French listening scores for executives

Executive	Pretest	Posttest	Executive	Pretest	Posttest
1	32	34	11	30	36
2	31	31	12	20	26
3	29	35	13	24	27
4	10	16	14	24	24
5	30	33	15	31	32
6	33	36	16	30	31
7	22	24	17	15	15
8	25	28	18	32	34
9	32	26	19	23	26
10	20	26	20	23	26

The Effects of Language Instruction

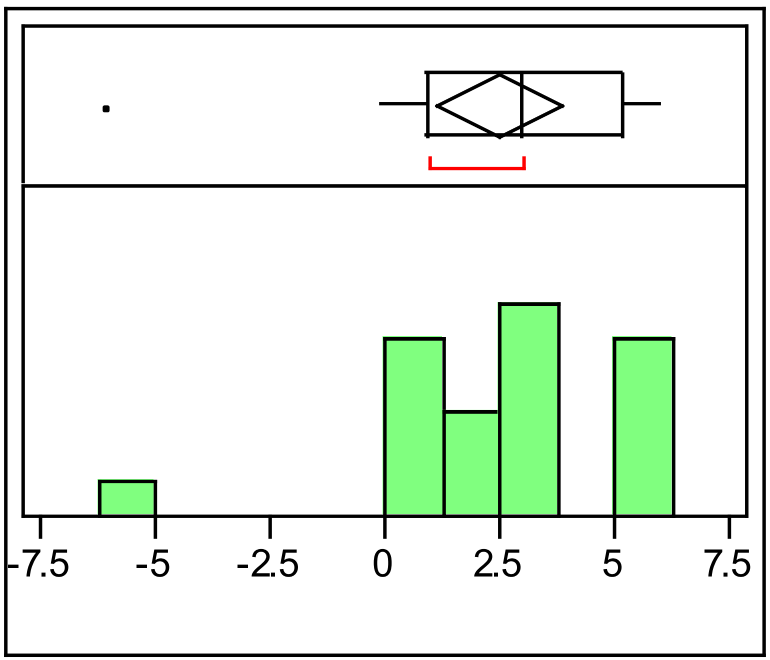
TABLE 7.2 French listening scores for executives

Executive	Pretest	Posttest	Gain	Executive	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	−6	19	23	26	3
10	20	26	6	20	23	26	3

How should we analyze these data?

Matched Pairs

- Y = Gain



Moments

Mean	2.5
Std Dev	2.8928223
Std Err Mean	0.6468547
upper 95% Mean	3.8538825
lower 95% Mean	1.1461175
N	20

The Effects of Language Instruction

H_0: $\mu = 0$ versus H_a: μ _____

Test Mean=value

Hypothesized Value	0
Actual Estimate	2.5
df	19
Std Dev	2.89282

	t Test
Test Statistic	3.8649
Prob > |t|	0.0010
Prob > t	0.0005
Prob < t	0.9995

Conclusion:

Example 2: Weight Change

A study was conducted to estimate the mean weight change of a female adult who quits smoking. The weights of eight female adults before they stopped smoking and five weeks after they stopped smoking were recorded. The differences, computed as "after -before," are given below.

Subject	1	2	3	4	5	6	7	8
After	154	181	151	120	131	130	121	128
Before	148	176	153	116	129	128	120	132
Difference	6	5	-2	4	2	2	1	-4

Here we have another example of a paired design.

(a) Compute the sample mean difference in weight.

(b) Compute the sample standard deviation of the differences.

Solution

(a) The sample mean difference is =1.75 pounds.

Note that the differences computed as “after - before” represent the weight gain for a subject.

A positive value indicates weight gain and a negative value indicates a weight loss.

(b) The sample standard deviation is SD =3.412 pounds

Paired *t*-Test

Hypotheses:

$H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$ or

$H_0 : \mu_D \leq 0$ versus $H_1 : \mu_D > 0$ or

$H_0 : \mu_D \geq 0$ versus $H_1 : \mu_D < 0$.

Data: The sample of n differences, generically written as $d_1, d_2, \ldots, d_n$ from which the sample mean difference $\bar{d}$ and the sample standard deviation of the differences s_D can be computed.

Test Statistic: $T = \dfrac{\bar{d} - 0}{s_D / \sqrt{n}}$ and the *null distribution* for the *T* variable is a $t_{(n-1)}$ *distribution*.

***p*-value:** We find the *p*-value for the test using the $t(n - 1)$ *distribution*.

The *direction of extreme* will depend on how the alternative hypothesis is expressed.

Independent Pairs

Frequently, the aim is to compare two populations. An analogous problem is to compare the effect of two "treatments" on the same population.

TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

The Standard Error for the Difference of Independent Means

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}$$

Example : A company that sells educational materials reports statistical studies to convince customers that its materials improve learning. One new product supplies "directed reading activities" for classroom use. These activities should improve the reading ability of elementary school pupils.

A consultant arranges for a group of 21 third grade students to take part in these activities for eight weeks. Another group of 23 students follows the same curriculum without the additional activities. The students are given a Degree of Reading Power at the end of the eight week period.

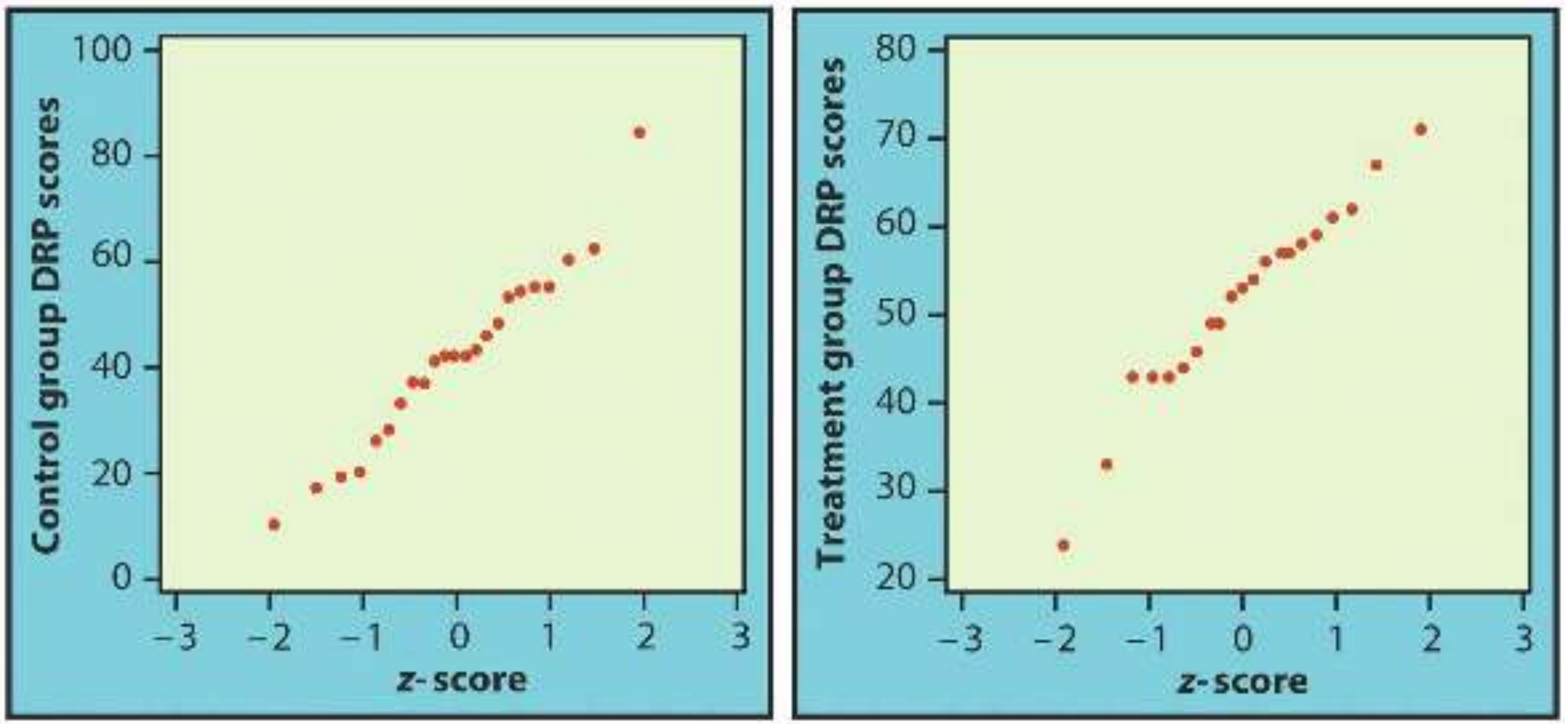
Are there any apparent differences between the groups???

Since we expect the activities to help, H_a should be one-sided.

TABLE 7.3 DRP scores for third-graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Normal probability plots for the two groups



Comparing Means

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent normally distributed samples. Then

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)$$

THE TWO-SAMPLE t CONFIDENCE INTERVAL

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1 and an independent SRS of size n_2 from another Normal population with unknown mean μ_2. The **confidence interval for $\boldsymbol{\mu_1 - \mu_2}$** given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least C no matter what the population standard deviations may be. The margin of error is

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The value of the degrees of freedom k is approximated by software or we use the smaller of $n_1 - 1$ and $n_2 - 1$.

Two-sample t test

- Two-sample t statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}$$

Distribution of T can be approximated by t distribution

- *Two-sided test:*

$H_0: \ \mu_X = \mu_Y$ against $H_a: \ \mu_X \neq \mu_Y$

reject H_0 if $|T| > t_{df,\alpha/2}$

- *One-sided test:*

$H_0: \ \mu_X = \mu_Y$ against $H_a: \ \mu_X > \mu_Y$

reject H_0 if $T > t_{df,\alpha}$

- *Degrees of freedom:*
    - Approximations for df provided by statistical software
    - *Satterthwaite* approximation

$$df = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_X^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_Y^2}{n}\right)^2}$$

      commonly used, conservative approximation

    - Otherwise: use $df = \min(m-1, n-1)$

Example: Effects of ozone

Study the effects of ozone by controlled randomized experiment

- 55 70-day-old rats were randomly assigned to two treatment or control
- *Treatment group:* 22 rats were kept in an environment containing ozone.
- *Control group:* 23 rats were kept in an ozone-free environment
- *Data:* Weight gains after 7 days

We are interested in the difference in weight gain between the treatment and control group.

Example: Effects of ozone

Data:

- Treatment group: $\bar{x} = 11.01$, $s_X = 19.02$, $m = 22$
- Control group: $\bar{x} = 22.43$, $s_X = 10.78$, $n = 23$

Testproblem:

- $H_0 : \mu_X = \mu_Y$ vs $H_a : \mu_X \neq \mu_Y$
- $\alpha = 0.05$, $df = \min(m-1, n-1) = 21$, $t_{21,0.025} = 2.08$

The value of the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{m}}} = \quad 2.46$$

The corresponding P-value is

$$\mathbb{P}(|T| \geq |t\ ) = \mathbb{P}(|T| \geq 2.46) = 0.023$$

Thus we reject the hypothesis that ozone has no effect on weight gain.

Example : A company that sells educational materials reports statistical studies to convince customers that its materials improve learning. One new product supplies "directed reading activities" for classroom use. These activities should improve the reading ability of elementary school pupils.

A consultant arranges for a group of 21 third grade students to take part in these activities for eight weeks. Another group of 23 students follows the same curriculum without the additional activities. The students are given a Degree of Reading Power at the end of the eight week period.

Are there any apparent differences between the groups???

Since we expect the activities to help, H_a should be one-sided.

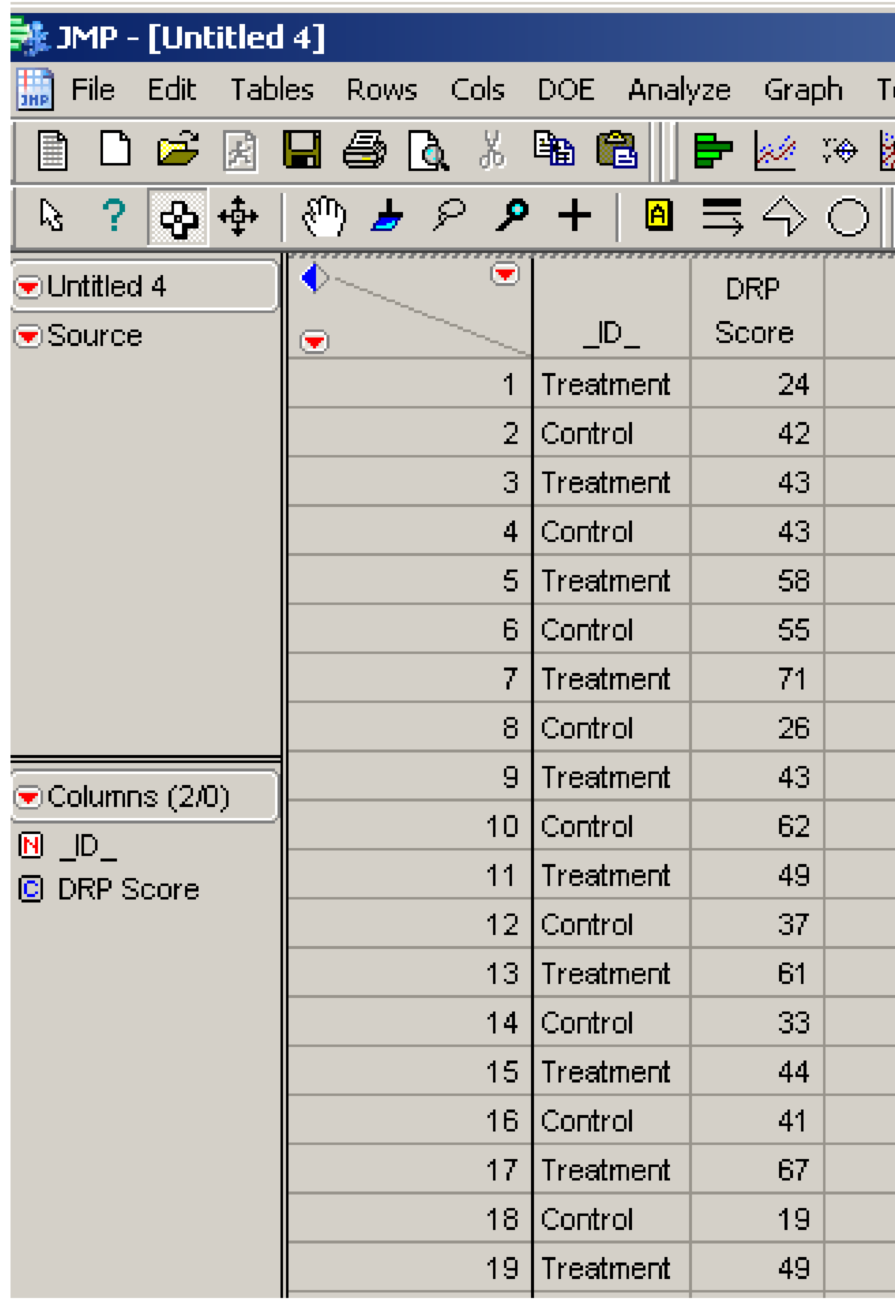
TABLE 7.3 DRP scores for third-graders

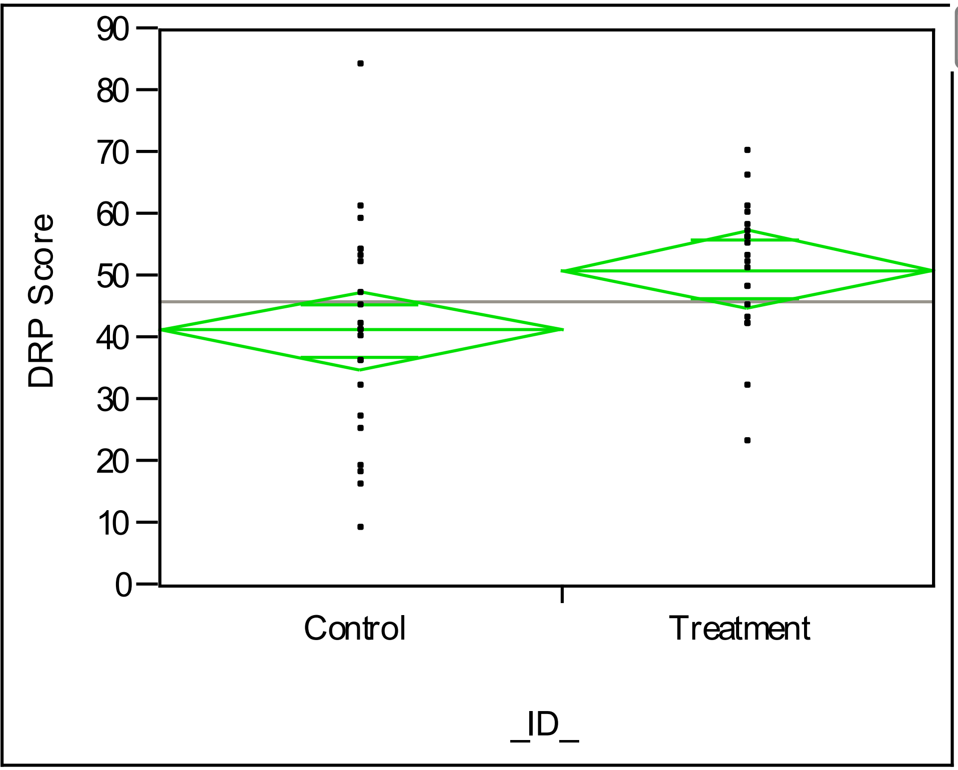
Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Analyze with Fit Y by X

Y =

X =





Oneway Anova

t Test

Assuming equal variances

	Difference	t Test	DF	Prob > |t|
Estimate	-9.9545	-2.267	42	0.0286
Std Error	4.3919			
Lower 95%	-18.8176			
Upper 95%	-1.0913			

UnEqual Variances

	Difference	t Test	DF	Prob > |t|
Estimate	-9.954	-2.311	37.8554	0.0264
Std Error	4.308			
Lower 95%	-18.846			
Upper 95%	-1.062			

We ask for a t-test. Use the one under "Unequal Variances."

Note: degrees of freedom = ________.

Example : We have information on healthy and failed companies, and we wish to determine if the mean ratio of current assets to current liabilities is different for the two groups. The data below provides information on 68 healthy and 33 failed firms.

TABLE 7.4 Ratio of current assets to current liabilities

Healthy firms						Failed firms		
1.50	0.10	1.76	1.14	1.84	2.21	0.82	0.89	1.31
2.08	1.43	0.68	3.15	1.24	2.03	0.05	0.83	0.90
2.23	2.50	2.02	1.44	1.39	1.64	1.68	0.99	0.62
0.89	0.23	1.20	2.16	1.80	1.87	0.91	0.52	1.45
1.91	1.67	1.87	1.21	2.05	1.06	1.16	1.32	1.17
0.93	2.17	2.61	3.05	1.52	1.93	0.42	0.48	0.93
1.95	2.61	1.11	0.95	0.96	2.25	0.88	1.10	0.23
2.73	1.56	2.73	0.90	2.12	1.42	1.11	0.19	0.13
1.62	1.76	2.22	2.80	1.85	0.96	2.03	0.51	1.12
1.71	1.02	2.50	1.55	1.69	1.64	0.92	0.26	1.15
1.03	1.80	0.67	2.44	2.30	2.21	0.13	0.88	0.09
1.96	1.81							