APPLIED STATISTICS

WEEK 7

Confidence Interval Estimation on Two Samples - PARTI

Inference about Two populations

* Confidence intervals for the difference between two means

$$LL < (\mu_1 - \mu_2) < UL$$

* Confidence Intervals for the difference between two proportions

$$LL < (P_1 - P_2) < UL$$

* Confidence intervals for the ratio of two variances

$$LL < \frac{\sigma_1^2}{\sigma_2^2} < UL$$

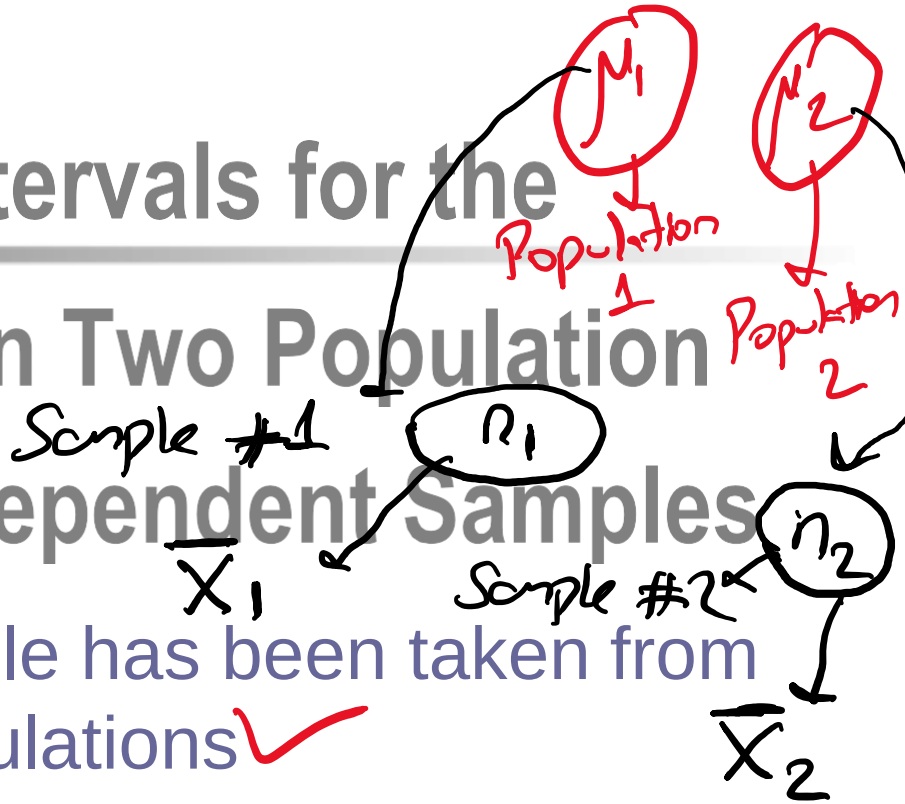
Estimating the Difference Between Two Means ($\mu_1 - \mu_2$)

Case I: Variances Known

Case II: Variances Unknown but Equal ($\sigma_1^2 = \sigma_2^2$)

Case III: Variances Unknown and Unequal ($\sigma_1^2 \neq \sigma_2^2$)

Confidence Intervals for the Difference between Two Population Means $\mu_1 - \mu_2$: Independent Samples



- Suppose a random sample has been taken from each of two different populations
- Suppose that the populations are independent of each other
- Then the random samples are independent of each other
- Then the sampling distribution of the difference in sample means is normally distributed

We use the statistic

$$\bar{X}_1 - \bar{X}_2$$

to compare two population means.

Sampling Distribution of the Difference of Two Sample Means

- Suppose population 1 has mean μ_1 and variance σ_1^2
  - From population 1, a random sample of size n_1 is selected which has mean $\bar{x}_1$ and variance s_1^2
- Suppose population 2 has mean μ_2 and variance σ_2^2
  - From population 2, a random sample of size n_2 is selected which has mean $\bar{x}_2$ and variance s_2^2
- Then the sample distribution of the difference of two sample means...

Sampling Distribution of the Difference of Two Sample Means

* Is normal, if each of the sampled populations is normal.
  * Approximately normal if the sample sizes n_1 and n_2 are large !!

* $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ ✓ (Mean)

* $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Standard deviation

Case 1: Confidence Interval on a Difference in Means, Variances Known

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variance σ_1^2 and σ_2^2, respectively, a **100(1 –α)% confidence interval for $\mu_1 - \mu_2$** is

Difference of Means
Population

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution

Lower Limit

Upper Limit

Question 1:

1. Confidence Interval for $\mu_1 - \mu_2$ (σ_1^2 and σ_2^2 KNOWN)

Case I (Voinces Known)

A study was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon, was measured. 50 experiments were conducted using engine type A and 75 experiments were done with engine type B. The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine A and 42 liters per gallon for engine B.

$n_1 = 50$ $n_2 = 75$

Find a 96% confidence interval on $\mu_B - \mu_A$; where μ_B and μ_A are population mean gas mileages for engines A and B, respectively. Assume that the population standard deviations are 6 and 8 for engines A and B, respectively.

$\sigma_1^2 = 36 \Leftarrow 6^2$

$\sigma_2^2 = 64 \leftarrow 8^2$

Case I: Constructing Confidence Interval for $\mu_B - \mu_A$ when Voinces are known

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

* The point estimate of $\mu_B - \mu_A$

CI → Point Estimate $\mp$ ME

$= \bar{X}_B - \bar{X}_A = 42 - 36 = 6.$

$(1-\alpha) = 0.96 \rightarrow \alpha = 0.04$

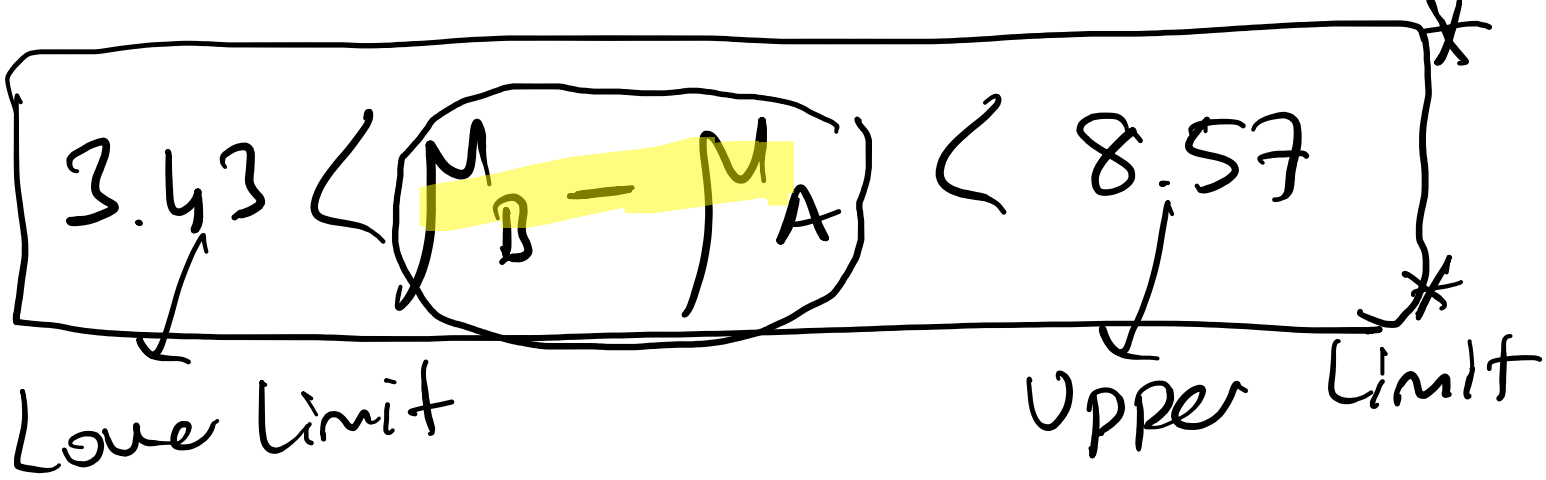
$\alpha/2 = 0.02$

$Z_{0.02} = 2.05$

$\alpha/2$

$$6 \pm 2.05 \sqrt{\frac{36}{50} + \frac{64}{75}}$$

$\Rightarrow (3.43, 8.57)$



Comment:

We are 96% confident that the interval from 3.43 to 8.57 contains the difference of the population means.

Question 2:

Aluminum Tensile Strength

Case I (Variances Known)

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10, \bar{x}_1 = 87.6, \sigma_1 = 1, n_2 = 12, \bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

Formula Case I

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645\sqrt{\frac{(1)^2}{10} + \frac{(1.5)^2}{12}} \le \mu_1 - \mu_2 \le 87.6 - 74.5 + 1.645\sqrt{\frac{(1^2)}{10} + \frac{(1.5)^2}{12}}$$

$$12.22 \le \mu_1 - \mu_2 \le 13.98$$

Comparing Two Population Means by Using Independent Samples: Variances Unknown

- Generally, the true values of the population variances σ_1^2 and σ_2^2 are not known.
- They have to be estimated from sample variances s_1^2 and s_2^2, respectively
- If the variances are not known and the two distributions involved are approximately normal, the t-dist. becomes involved, as in the case of a single sample.

When two Population Variances are Unknown

Equal Variances (Case 2)

Unequal Variances (Case 3)

Case 2: Comparing Two Population Means by Using Independent Samples: Variances Unknown but Equal

CI → Point Estimate ± ME

Case2: Confidence Interval on the Difference in Means, Variance Unknown

Assumption: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$df = n_1 + n_2 - 2$

$t_{\alpha/2, df}$

If $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 are the sample means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown but equal variances, then a 100(1 - α)% confidence interval on the difference in means $\mu_1 - \mu_2$ is

Pooled Estimated of the Common population standard deviation

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Critical t value at $\alpha/2$ with $df = n_1 + n_2 - 2$

where $s_p = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)}$ is the pooled estimate of the common population standard deviation, and $t_{\alpha/2,\, n_1+n_2-2}$ is the upper α/2 percentage point of the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom.

Case II: Variances Unknown but Equal

$$\sigma_1 = \sigma_2 = \sigma$$

$$S_p^2 = \frac{(n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2}{n_1+n_2-2}$$

& Pooled estimate of variance

(Pooled Estimator for grouping standard deviations)

Question 3: (Variances Unknown but Equal)

Case 2

Cement Hydration

Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$ with a sample standard deviation of $s_1 = 5.0$, and 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$ with a sample standard deviation of $s_2 = 4.0$. Assume that weight percent calcium is normally distributed with same standard deviation. Find a 95% confidence interval on the difference in means, $\mu_1 - \mu_2$, for the two types of cement.

$\bar{X}_1 = 90, \; S_1 = 5, \; n_1 = 10$ → Type I Cement

$\bar{X}_2 = 87, \; S_2 = 4, \; n_2 = 15$ → Type II Cement

S_p^2 = Pooled estimate of variance

$$S_p^2 = \frac{9(5)^2 + 14(4)^2}{10 + 15 - 2} = 19.52$$

$$S_p = \sqrt{19.52} = 4.4$$

Pooled estimate of the common standard deviation

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.025,23}\, Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$df = n_1 + n_2 - 2 = 10 + 15 - 2 = 23$$

$$90 - 87 \pm 2.069\ (4.4)\sqrt{\frac{1}{10}+\frac{1}{15}}$$

$$-0.72 < \mu_1 - \mu_2 < 6.72$$

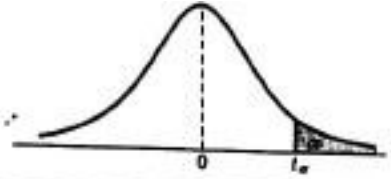
Lower Limit

Lower Limit

We are 95 % confident that the interval from −0.72 to 6.72 contains the difference of the population means

$\nu = df$

Table A.4 Critical Values of the t-Distribution



α

ν	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056

$\alpha/2 = 0.025$

2.069

Question 4: → Case 2

The article "**Microinvertebrate Community Structure as an Indicator of Acid Mine Pollution**," published in the *Journal of Environmental Pollution*, reports on an investigation undertaken in Cane Creek, Alabama, to determine the relationship between selected physiochemical parameters and different measures of macroinvertebrate community structure. One facet of the investigation was an evaluation of the effectiveness of a numerical species diversity index to indicate aquatic degradation due to acid mine drainage. Conceptually, a high index of macroinvertebrate species diversity should indicate an unstressed aquatic system, while a low diversity index should indicate a stressed aquatic system.

Consider that two independent sampling stations were chosen for a study, one located downstream from the acid mine discharge point and the other located upstream. Foe 12 monthly samples collected at the downstream station, the species diversity index had a mean value of $\bar{x}_1 = 3.11$ and a standard deviation $s_1 = 0.771$, while 10 monthly samples collected at the upstream station had a mean index value $\bar{x}_2 = 2.04$ and a standard deviation $s_2 = 0.448$.

Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

Question 4:

Let μ_1 and μ_2 represent the population means, respectively, for the species diversity indices at the downstream and upstream stations. We wish to find a 90% confidence interval for $\mu_1 - \mu_2$. Our point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 3.11 - 2.04 = 1.07.$$

The pooled estimate, s_p^2, of the common variance, σ^2, is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)(0.771^2) + (9)(0.448^2)}{12 + 10 - 2} = 0.417.$$

Taking the square root, we obtain $s_p = 0.646$. Using $\alpha = 0.1$, we find that $t_{0.05} = 1.725$ for $v = n_1 + n_2 - 2 = 20$ degrees of freedom. Therefore, the 90% confidence interval for $\mu_1 - \mu_2$ is

$$1.07 - (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}},$$

which simplifies to $0.593 < \mu_1 - \mu_2 < 1.547.$

Case3: Confidence Interval on the Difference in Means, Variance Unknown and Unequal

Assumption : $\sigma_1^2 \neq \sigma_2^2$ They are unequal

If $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate 100(1 - α)% confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$t_{\alpha/2,v}$ the upper α /2 percentage point of the t distribution with v degrees of freedom.

$t_{\alpha/2,v}$

Comparing Population Means with Unknown **AND** Unequal Population Standard Deviations

Use the formula for the *t*-statistic shown if it is not reasonable to assume the population standard deviations are equal.

TEST STATISTIC FOR NO DIFFERENCE IN MEANS, UNEQUAL VARIANCES

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad [11\text{–}5]$$

The degrees of freedom are adjusted downward by a rather complex approximation formula. The effect is to reduce the number of degrees of freedom in the test, which will require a larger value of the test statistic to reject the null hypothesis.

DEGREES OF FREEDOM FOR UNEQUAL VARIANCE TEST

$$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad [11\text{–}6]$$

ROUND DOWN!

Question 5: (Case 3: Unknown and Unequal Case)

A study was conducted by the department of Zoology at the Virginia Tech to estimate the difference in the amounts of the chemical orthophosphorus measured at two different locations on the James River. Orthophosphorus was measured in milligrams per liter. 15 samples were collected from station 1, and 12 samples were obtained from station 2. The 15 samples from station 1 had an average orthophosphorus content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average content of 1.49 milligrams per liter and a standard deviation of 0.80 milligram per liter.

Find a 95% confidence interval for the difference in the true average orthophosphorus contents at these two stations, assuming that the observations came from normal populations with different variances.

Station 1: $n_1 = 15$, $\bar{X}_1 = 3.84$, $S_1 = 3.07$

Station 2: $n_2 = 12$, $\bar{X}_2 = 1.49$, $S_2 = 0.8$

$$\nu = \frac{\left[(3.07)^2/15 + (0.80)^2/12\right]^2}{\frac{\left((3.07)^2/15\right)^2}{14} + \frac{\left((0.80)^2/12\right)^2}{11}}$$

$$\nu = 16.3 \approx \boxed{16}$$

$$\bar{X}_1 - \bar{X}_2 = 3.84 - 1.49 = 2.35$$

$$t_{0.025,16} = 2.120$$

$$2.35 \pm 2.120 \sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}}$$

$$0.60 < \mu_1 - \mu_2 < 4.10$$

Hence, we are 95% confident that the interval from 0.60 to 4.10 milligrams per liter contains the difference of the true average orthophosphorus contents for these two locations (stations).

Table A.4 Critical Values of the t-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064

2.120