Chapter 9 One- and Two-Sample Estimation Problems

吳明龍 副教授
成大資訊系/醫資所
minglong.wu@csie.ncku.edu.tw
辦公室:資訊系館 12 樓

9.2 Statistical Inference

- Statistical inference
  - Classical method: inferences are based strictly on information obtained from a random sample selected from the population.
  - Bayesian method: utilizes prior subjective knowledge about the probability distribution of the unknown parameters in conjunction with the information provided by the sample data. (Chapter 18)
  - This chapter use classical methods.
- Statistical inference may be divided into two major areas:
  - Estimation
  - Tests of hypotheses

9.3 Classical Methods of Estimation

- A point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$
  - The value $\bar{x}$ of the statistic $\bar{X}$, is a point estimate of the population parameter μ.
  - $\hat{p} = x/n$ is a point estimate of the true proportion *p* for a binomial experiment.
- Definition 9.1: A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Classical Methods of Estimation

- Example 9.1: Show that S^2 is an unbiased estimator of the parameter σ^2.
- This example illustrates why we divide by n-1 rather than n when the variance is estimated.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Now

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma_{X_i}^2 - n\sigma_{\overline{X}}^2\right).$$

However,

$$\sigma_{X_i}^2 = \sigma^2 \quad \text{for } i = 1, 2, \cdots, n \quad \text{and} \quad \sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}.$$

Therefore,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Classical Methods of Estimation

- Definition 9.2: If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the most efficient estimator of θ.
  - $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased (centered at θ).
  - $\hat{\Theta}_1$ has a smaller variance than $\hat{\Theta}_2$ and is therefore more efficient.

    $\therefore$ the estimator of $\theta : \hat{\Theta}_1$

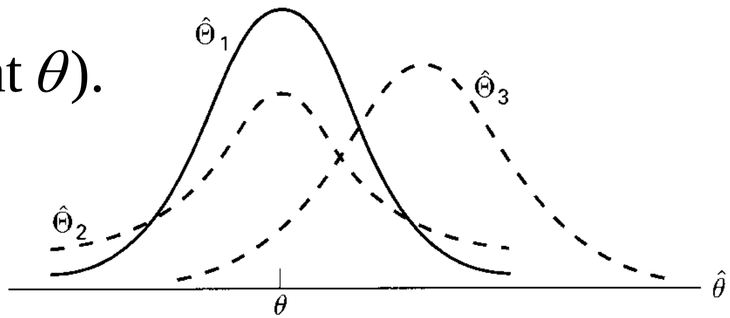

Figure 9.1 Sampling distributions of different estimators of θ.

$\hat{\Theta}$

Interval Estimation

- Unlikely to estimate the population parameter exactly.
- Accuracy increases with large samples.
- In many situations, preferable to determine an interval within which we would expect to find the value of the parameter.
- Such an interval is called an interval estimate.
  - $\hat{\theta}_L < \theta < \hat{\theta}_U$
- As the sample size increases, we know that $\sigma^2_{\bar{X}} = \sigma^2 / n$ decreases, and consequently our estimate is likely to be closer to the parameter μ, resulting in a shorter interval.
- An interval estimate might be more informative.

Interpretation of Interval Estimation

- $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha$
- We have a probability of 1 - α of selecting a random sample that will produce an interval containing θ.
  - The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$ is called a (1 - α)100% confidence interval.
  - 1 - α is called confidence degree.
  - $\hat{\theta}_L$ and $\hat{\theta}_U$ are called the lower and upper confidence limits.
- We prefer <u>a short interval with a high degree of confidence</u>.

9.4 Single Sample: Estimating the Mean

- According to the central limit theorem, we can expect the sampling distribution of $\overline{X}$ to be approximately normally distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma / \sqrt{n}$.

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

$$P(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

$$P(\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

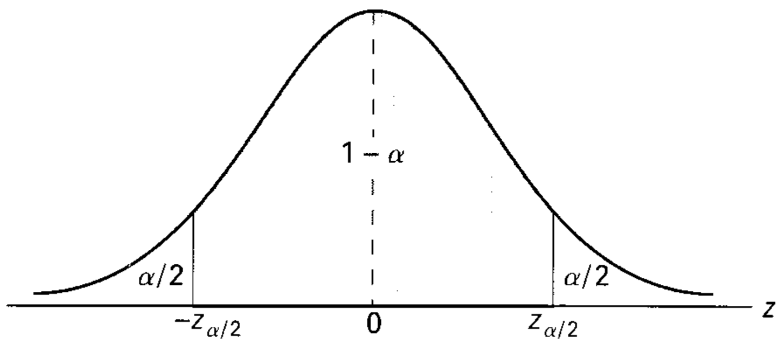

Figure 9.2 $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

- Confidence interval of μ: $\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Single Sample: Estimating the Mean

- Different samples will yield different values of $\overline{X}$ and therefore produce different interval estimates of the parameter μ.

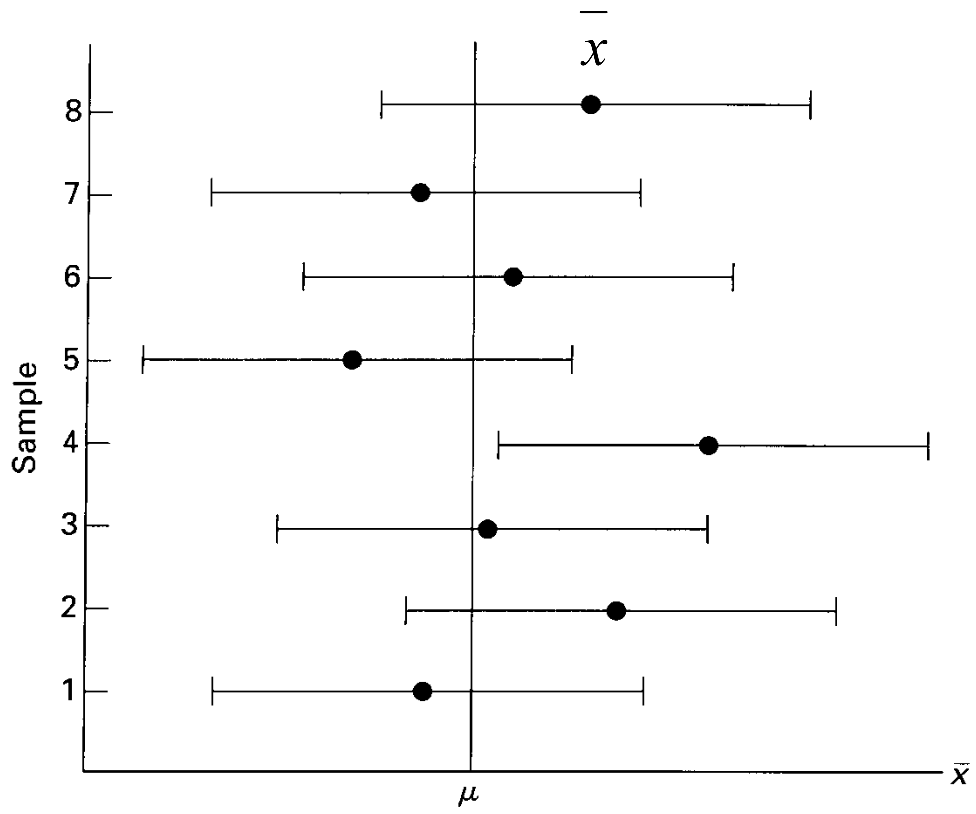

Figure 9.3 Interval estimates of μ for different samples.

Single Sample: Estimating the Mean

- Ex9.2: The average zinc concentration recovered from a sample of zinc measurements in 36 different locations is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3.
  - Solution

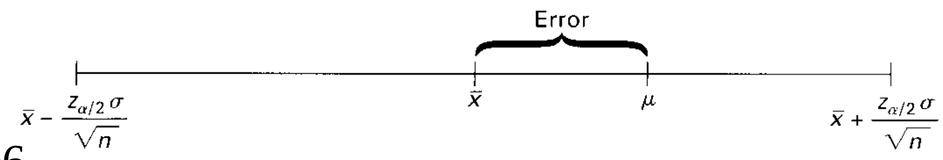

Figure 9.4 Error in estimating μ by $\bar{x}$.

For 95% confidence interval $\Rightarrow z_{0.025} = 1.96$

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right) \Rightarrow 2.50 < \mu < 2.70$$

For 99% confidence interval $\Rightarrow z_{0.005} = 2.575$

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right) \Rightarrow 2.47 < \mu < 2.73$$

TABLE A.3 (continued) Areas Under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993

Single Sample: Estimating the Mean

- Theorem 9.1: If $\bar{x}$ is used as an estimate of μ, we can then be (1 - α)100% confident that the error will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- Theorem 9.2: If $\bar{x}$ is used as an estimate of μ, we can then be (1 - α)100% confident that the error will not exceed a specified amount ***e*** when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

- Example 9.3: How large a sample is required in Example 9.2 if we want to be 95% confident that our estimate of μ is off by less than 0.05?
  - Solution

$$n = \left(\frac{(1.96)(0.3)}{0.05}\right)^2 = 138.3 \therefore \text{sample size} = 139.$$

Estimating the Mean with σ Unknown

- If we have a random sample from a normal distribution, then the random variable *T* has a student's *t*-distribution with *n* – 1 degrees of freedom.

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

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

$$P(-t_{\alpha/2} < \frac{\overline{X} - \mu}{S / \sqrt{n}} < t_{\alpha/2}) = 1 - \alpha$$

$$P(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

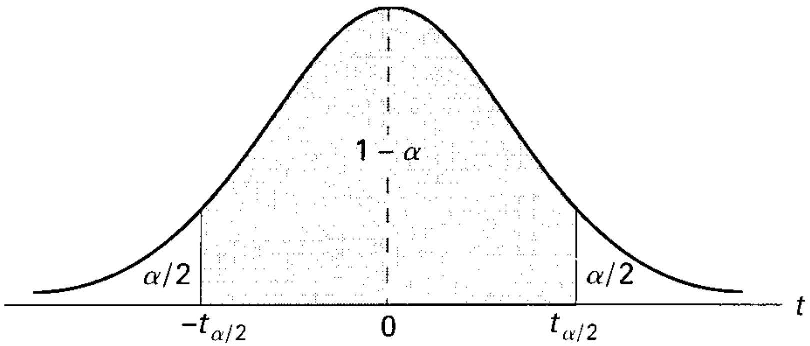

Figure 9.5 $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha.$

- Confidence interval of μ: $\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$

Estimating the Mean with σ Unknown

Example 9.5: The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean of all such containers, assuming an approximate normal distribution.

Solution The sample mean and standard deviation for the given data are

$$\bar{x} = 10.0 \quad \text{and} \quad s = 0.283.$$

Using Table A.4, we find $t_{0.025} = 2.447$ for $v = 6$ degrees of freedom. Hence the 95% confidence interval for μ is

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

which reduces to $9.74 < \mu < 10.26$.

t-Distribution

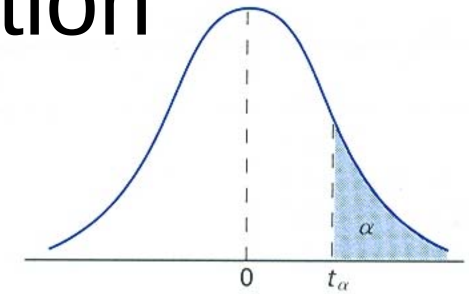

TABLE A.4 Critical Values of the t-Distribution

	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	0.537	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

Large-Sample Confidence Interval

- Large-sample confidence interval: when normality cannot be assumed, σ is unknown, and $n \geq 30$, s can replace σ and the confidence interval

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

may be used.

  – *s* will be very close to the true σ and thus the central limit theorem prevails.

9.5 Standard Error of a Point Estimate

The variance of $\overline{X}$: $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

Thus the standard deviation of $\overline{X}$ or standard error of $\overline{X}$: $\sigma/\sqrt{\mathrm{n}}$.

Confidence limit: $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2}\mathrm{s.e.}(\bar{x})$.

Confidence limit with σ unknown: $\bar{x} \pm t_{\alpha/2}\dfrac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2}\mathrm{s.e.}(\bar{x})$.

- Width of confidence intervals become shorter as the quality of the corresponding point estimate becomes better.

9.6 Prediction Interval

- Sometimes, some experimenters may also be interested in predicting the possible value of a future observation.
- Some customers may require a statement regarding the uncertainty of one single observation.
- The type of requirement is nicely fulfilled by the construction of a prediction interval.
- Assume: a natural point estimator of a new observation is $\overline{X}$, and the variance of $\overline{X}$: σ^2/n
- The development of a prediction interval is displayed by beginning with a normal random variable x_0 - $\overline{x}$

x_0 : the new observation, the variance is σ^2

$\overline{x}$: comes from the sample

$$z = \frac{x_0 - \overline{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \overline{x}}{\sigma\sqrt{1+1/n}}$$

Using $\Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha \Rightarrow \overline{x} - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}} < x_0 < \overline{x} + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}$

Prediction Interval

- For a normal distribution of measurements with unknown mean μ and known variance σ^2, a (1 - α)100% prediction interval of a future observation, x_0, is $\bar{x} - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}$

  where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Prediction Interval

- Example 9.7: Due to the decreasing of interest rates, the First Citizens Bank received a lot of mortgage applications. A recent sample of 50 mortgage loans resulted in an average of $257,300. Assume a population standard deviation of $25,000. If a next customer called in for a mortgage loan application, find a 95% prediction interval on this customer's loan amount.

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1+\tfrac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1+\tfrac{1}{n}}$$

$$\bar{x} = \$257{,}300, \quad z_{0.025} = 1.96$$

$$257300 - 1.96 \cdot 25000\sqrt{1+\tfrac{1}{50}} < x_0 < 257300 + 1.96 \cdot 25000\sqrt{1+\tfrac{1}{50}}$$

$$\therefore \text{prediction interval } (\$207{,}812.43, \$306{,}787.57)$$

- Prediction interval provides a good estimate of the location of a future observation.
- The estimation of future observation is quite different from the estimation of sample mean.

Prediction Interval

- For a normal distribution of measurements with unknown mean μ and unknown variance σ^2, a (1 - α)100% prediction interval of a future observation, x_0, is $\bar{x} - t_{\alpha/2} s\sqrt{1+\frac{1}{n}} < x_0 < \bar{x} + t_{\alpha/2} s\sqrt{1+\frac{1}{n}}$,

  where $t_{\alpha/2}$ is the t-value with $v = n$ -1 degree-of-freedom, leaving an area of $\alpha/2$ to the right.

Prediction Interval

- Example 9.8: A meat inspector has randomly measured 30 packs of acclaimed 95% lean beef. The sample resulted in the mean 96.2% with the sample standard deviation of 0.8%. Find a 99% prediction interval for a new pack. Assume normality.

$$v = 29, t_{0.005} = 2.756$$

$$96.2 - 2.756 \cdot 0.8\sqrt{1+\tfrac{1}{30}} < x < 96.2 + 2.756 \cdot 0.8\sqrt{1+\tfrac{1}{30}}$$

$$\therefore \text{prediction interval}\ (93.96, 98.44)$$

- An observation is an outlier if it falls outside the prediction interval computed without inclusion of the questionable observation in the sample.

Tolerance Limits

- Tolerance limits: For a normal distribution of measurements with unknown mean μ and unknown variance σ^2, tolerance limits are given by $\bar{x} \pm ks$, where k is determined so that one can assert with 100(1 - γ)% confidence that the given limits contain at least the proportion (1 - α) of the measurements.
- Example 9.8: A machine is producing metal pieces that are cylindrical in shape. A sample of these pieces is taken and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Find the 99% tolerance limits that will contain 95% of the metal pieces produced by this machine, assuming an approximate normal distribution.
  – Solution $\bar{x} = 1.0056,\ s = 0.0246$

    $\text{Look up Table A.7}, n = 9, 1 - \gamma = 0.99, \text{and } 1\text{-}\alpha = 0.95 \Rightarrow k = 4.550$

    $\therefore 99\% \text{ tolerance limits}: \ 1.0056 \pm 4.550 \cdot 0.0246$

TABLE A.7* Tolerance Factors for Normal Distributions

	$\gamma = 0.05$				$\gamma = 0.01$		
	$1-\alpha$				$1-\alpha$		
n	0.90	0.95	0.99	n	0.90	0.95	0.99
2	32.019	37.674	48.430	2	160.193	188.491	242.300
3	8.380	9.916	12.861	3	18.930	22.401	29.055
4	5.369	6.370	8.299	4	9.398	11.150	14.527
5	4.275	5.079	6.634	5	6.612	7.855	10.260
6	3.712	4.414	5.775	6	5.337	6.345	8.301
7	3.369	4.007	5.248	7	4.613	5.488	7.187
8	3.136	3.732	4.891	8	4.147	4.936	6.468
9	2.967	3.532	4.631	9	3.822	4.550	5.966
10	2.839	3.379	4.433	10	3.582	4.265	5.594
11	2.737	3.259	4.277	11	3.397	4.045	5.308
12	2.655	3.162	4.150	12	3.250	3.870	5.079
13	2.587	3.081	4.044	13	3.130	3.727	4.893
14	2.529	3.012	3.955	14	3.029	3.608	4.737
15	2.480	2.954	3.878	15	2.945	3.507	4.605
16	2.437	2.903	3.812	16	2.872	3.421	4.492
17	2.400	2.858	3.754	17	9.808	3.345	4.393
18	2.366	2.819	3.702	18	2.753	3.279	4.307
19	2.337	2.784	3.656	19	2.703	3.221	4.230
20	2.310	2.752	3.615	20	2.659	3.168	4.161
25	2.208	2.631	3.457	25	2.494	2.972	3.904
30	2.140	9.549	3.350	30	2.385	2.841	3.733
35	2.090	2.490	3.272	35	2.306	2.748	3.611
40	2.052	2.445	3.213	40	2.247	2.677	3.518

*Adapted from C. Eisenhart, M. W. Hastay, and W. A. Wallis, *Techniques of Statistical Analysis*, Chapter 2, McGraw-Hill Book Company, New York, 1947. Used with permission of McGraw-Hill Book Company.

Distinction Among Confidence Intervals, Prediction Intervals, and Tolerance Intervals

- Confidence intervals: population mean
- Tolerance limits: a tolerance interval is generally longer than a confidence interval with the same degree of confidence.
- Prediction limits: determine a bound of a future observation value.