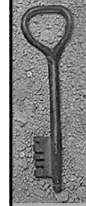


Statistical Inference: Point and Interval Estimation


GOG 502/PLN 504 Youqin Huang



Statistical Inference

- ◆ Principles of parameter estimation
- ◆ Estimation methods
- ◆ Point estimation
- ◆ Confidence interval
 - For the mean
 - For proportion
 - t-distribution for small sample


GOG 502/PLN 504 Youqin Huang



Statistical Inference

- ◆ Sample statistics \rightarrow population parameters
- ◆ Two tasks are intertwined
 - Estimating the parameters
 - How good are the estimates?
- ◆ The fundamental use of normal probability function
 - Central Limit Theorem
 - Most estimating methods assume normal distribution
 - Many variables follow or can be transformed into normal distribution


GOG 502/PLN 504 Youqin Huang



Principles of Parameter Estimation

- ◆ Unbiased
 - The expected value of the estimate is equal to population parameter
- ◆ Consistent
 - As n (sample size) approaches N (population size), estimator converges to the population parameter
- ◆ Efficient
 - With the smallest variance.
- ◆ Sufficient
 - Contains all information about the parameter through a sample of size n


GOG 502/PLN 504 Youqin Huang



Estimation Methods

- ◆ Assume probability distribution?
 - Parametric, nonparametric
- ◆ Common approaches
 - Least squares (LS)
 - Minimize the sum of squares of the deviation
 - Used in linear regression
 - Maximum likelihood estimation (MLE)
 - Estimate parameter that is most consistent with the data
 - Widely used
 - Minimum Chi-squared
 - frequencies


GOG 502/PLN 504 Youqin Huang



Estimation: Point vs. Interval

- ◆ Point estimation
 - Use one single number as the best estimator for a specific population parameter
 - “Point estimator” (e.g. estimator for mean=15)
- ◆ Interval estimation, “Confidence Interval”
 - use a range of numbers within which the parameter is believed to fall (lower bound, upper bound)
 - e.g. (10, 20)

GOG 502/PLN 504 Youqin Huang



Point Estimation

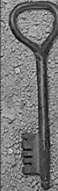
- ◆ Sample mean and std. dev. are the most popular point estimates of population mean and std. dev.

$$\hat{\mu} = \bar{Y} = \frac{\sum Y_i}{n}$$

$$\hat{\sigma} = s = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

- Why is sample mean a good estimator?
- Why n-1?

GOG 502/PLN 504 Youqin Huang



Point Estimation

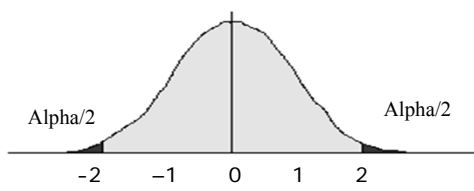
- ◆ Point estimation is rarely 100% accurate
- ◆ Accuracy depends on the characteristics of sampling distribution (z, t, chi-square, F distribution)
- ◆ Knowing the shape of the sampling distribution, you can figure out the general range of error that a given point estimate might miss by, or the probability of a point estimate that is true

GOG 502/PLN 504 Youqin Huang

Interval Estimation

- ◆ Confidence Interval
 - “A range of values around a point estimate that makes it possible to state the probability that an interval contains the population parameter between its lower and upper bounds.”
- ◆ It involves a range and a probability
 - **Confidence coefficient**: the probability that the interval contains the parameter, e.g. 0.90, 0.95, 0.99.
 - **Error probability (α)**: the probability that an interval does not contain the parameter.
 - $\alpha=1$ -confidence coefficient
- ◆ Examples:
 - We are 95% confident that the mean number of CDs owned by grad students is between 20 and 45

GOG 502/PLN 504 Youqin Huang



- ◆ Upper, lower bound?
- ◆ Confidence coefficient?
- ◆ Error probability?

GOG 502/PLN 504 Youqin Huang

Confidence Interval for Mean

- ◆ CLT states that sampling distribution of \bar{Y} is approximately normal.
 - With probability 0.95, \bar{Y} falls within $1.96\sigma_{\bar{Y}}$ units of the parameter μ .
 - Once the sample is selected, with probability 0.95 a \bar{Y} value occurs such that the interval $\bar{Y} \pm 1.96\sigma_{\bar{Y}}$ contains the population mean μ .
- ◆ For $n \geq 30$, a good approximation for $\sigma_{\bar{Y}}$ is sample standard deviation s (“standard error”).

$$\hat{\sigma}_{\bar{Y}} = \frac{s_{\bar{Y}}}{\sqrt{n}}$$

GOG 502/PLN 504 Youqin Huang

Example: 95% Confidence Interval

- ◆ Suppose a sample of 100 students with mean SAT score of 1020, standard deviation of 200
 - How do we find the 95% Confidence Interval for the population mean?
- ◆ We know that:
 - 1. The sampling distribution is normally distributed
 - 2. Therefore 95% of samples will yield a mean estimate within 2 standard deviations of the population mean (μ)
- ◆ Thus, 95% of estimates we make are within two “standard errors” of \bar{Y}

GOG 502/PLN 504 Youqin Huang

95% Confidence Interval

$$95\% \text{ CI: } \bar{Y} \pm 2(\hat{\sigma}_Y) \quad \hat{\sigma}_Y = \frac{s_Y}{\sqrt{n}}$$

$$\begin{aligned} \bar{Y} \pm 2\left(\frac{s}{\sqrt{n}}\right) &= 1020 \pm (2)\left(\frac{200}{\sqrt{100}}\right) \\ &= 1020 \pm 2 \times 20 \\ &= 1020 \pm 40 \end{aligned}$$

95% CI: (980, 1060)

GOG 502/PLN 504 Youqin Huang

Confidence Interval for Mean

◆ Similarly,

$$68\% \text{ CI: } \bar{Y} \pm 1\left(\frac{s}{\sqrt{n}}\right) = 1020 \pm 20$$

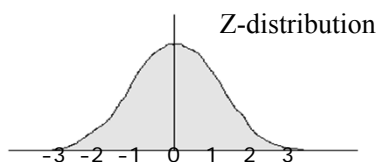
$$99\% \text{ CI: } \bar{Y} \pm 3\left(\frac{s}{\sqrt{n}}\right) = 1020 \pm 3 \times 20$$

- ◆ Q: Which one is a larger range?
- ◆ The larger the range, the more likely that the true mean will fall in it
 - It is a safer bet if you specify a very wide range

GOG 502/PLN 504 Youqin Huang

Confidence Interval for Mean

- ◆ What if we want to know the confidence interval for any number, other than 68%, 95%, 99%?
- Answer: the “Z-distribution” (standard normal), only if the distribution of interest is normal distribution (large n)



The Z-Distribution

- ◆ Recall:
 - 1. Z-score corresponds to # of standard deviations from mean
 - 2. Probability is equal to area under a curve
- ◆ Statisticians have calculated the exact area, thus probability under the curve associated with every Z-value (Table in appendix)


GOG 502/PLN 504 Youqin Huang

The Z-Distribution

- ◆ We already know area for with some Z-values:
 - Area between -1 and $+1$ Z score is roughly .68
 - Area between -2 and $+2$ Z score is roughly .95
 - Area between -3 and $+3$ Z score is roughly .99
- ◆ Note: we are typically interested in area from $-Z$ to $+Z$, or beyond $-Z$ and $+Z$

GOG 502/PLN 504 Youqin Huang

TABLE A Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z , probabilities are found by symmetry)

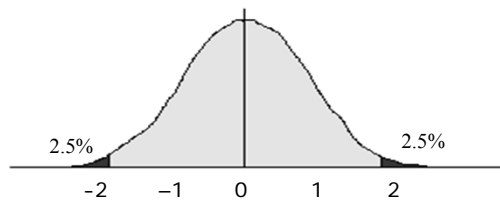


z	Second Decimal Place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0066	.0064	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013									
3.5	.00023									
4.0	.000017									
4.5	.00000049									
5.0	.000000287									

Source: R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1963).

The Z-Distribution

- ◆ As we know, Z of 1.96 corresponds to 95% of area ($p=.95$)
 - Cutting off 5% on the tails... 2.5% on each side




GOG 502/PLN 504 Youqin Huang

Confidence Interval for Mean

- ◆ General formula for CI:

$$C.I.: \bar{Y} \pm Z_{\alpha/2} (\sigma_{\bar{Y}})$$
 - \bar{Y} -bar is the sample mean
 - σ is the standard error of the mean (σ/\sqrt{n})
 - Z is the critical Z-value for a given level of confidence
- ◆ C. I. increases as the confidence coefficient increases
- ◆ C. I. decreases as the sample size increases


GOG 502/PLN 504 Youqin Huang



Confidence Interval for Proportion

- ◆ Nominal/ordinal data, or continuous data but measured in categories, e.g. income 0-10,000, 10,001-20,000, 20,001-30,000...
- ◆ Summary parameter: proportion
 - Recall: proportion is a type of mean
- ◆ Widely used in social sciences:
 - % of foreign born who attended college
 - % of hh with income below poverty level
 - % of hh who own their homes
 - ...

GOG 502/PLN 504 Youqin Huang




Confidence Interval for Proportion

- ◆ $P(y=1)=\pi$, $P(y=0)=1-\pi$
- ◆ π is the population proportion (similar to μ), $\hat{\pi}$ is sample proportion
- ◆ Standard error

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$
- ◆ Confidence interval:

$$\hat{\pi} \pm Z_{\alpha/2}(\hat{\sigma}_{\hat{\pi}})$$

GOG 502/PLN 504 Youqin Huang



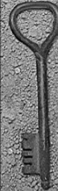
C.I. for Proportion: Example

- ◆ 1994 GSS on attitude on legal abortion. Out of 1934 people, 895 said “yes”, 1039 said “no”. Estimate the proportion of U.S. population that would say “yes” to this question
- ◆ $N=1934$, $\hat{\pi} = 895/1934=0.46$

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.46(1-0.46)}{1934}} = 0.011$$
- ◆ Confidence interval (95%):

$$\hat{\pi} \pm Z_{\alpha/2}(\hat{\sigma}_{\hat{\pi}}) = 0.46 \pm 1.96 * 0.011$$
 - So, 95% C.I. is (0.44, 0.48)

GOG 502/PLN 504 Youqin Huang



Confidence Intervals: Small n

- ◆ The formula for C.I. works with the assumption of large n.
 - we assume the sampling distribution is normal
 - Z-distribution probability, Z-score
- ◆ What if n is not large (n<30)?
 - The C.L.T. holds only if n is “large”
 - we can’t use Z-scores to determine probabilities under the normal curve

GOG 502/PLN 504 Youqin Huang

Confidence Intervals: Small n

◆ Solution: Use the “t statistic”, “t-distribution”

– t statistic

$$t = \frac{\bar{Y} - \mu}{\hat{\sigma}_Y} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

– Also called *Student's t*.

◆ T-distribution

- The sampling distribution of the t statistic with $n-1$ degree of freedom
- It is actually a *set* of distributions; each is relevant to a different sample size, $df=n-1$
- The spread of t-distribution depends on the d.f.
- When $n>30$, identical to z-distribution

GOG 502/PLN 504 Youqin Huang

CI for small n: t Distribution

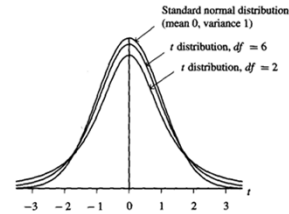
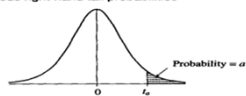


Figure 6.9 t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (df) increase, and the two distributions are practically identical when $df > 30$.

GOG 502/PLN 504 Youqin Huang

TABLE B t Distribution. Values corresponding to various right-hand tail probabilities



df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.7063	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.608
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Source: "Table of Percentage Points of the t-Distribution."

Confidence Interval: Small n

◆ Small sample C. I.:

$$\text{C.I.: } \bar{Y} \pm t_{\alpha/2}(\hat{\sigma}_Y)$$

◆ Again, the standard error can be estimated using the sample standard deviation:

$$\text{C.I.: } \bar{Y} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

GOG 502/PLN 504 Youqin Huang



Summary

- ◆ Principles of parameter estimation
- ◆ Point vs. interval estimation
- ◆ Point estimation of mean, s.d.
- ◆ Interval estimation: Confidence Interval
 - Mean
 - Proportion
 - Issue of sample size
 - Small n, t-distribution

GOG 502/PLN 504 Youqin Huang