W1211 Introduction to Statistics
Lecture 22

Wei Wang

Nov 20, 2012

What we talked about last lecture

- Interpretation of Confidence Intervals

What we talked about last lecture

- Interpretation of Confidence Intervals
- Construct a Confidence Interval for a normal population mean μ when the variance σ^2 is assumed known.

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Interval for the Mean of a Normal Population when Variance is assumed known

- A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

- $z_{\alpha/2}$ is the upper $(100 \cdot \alpha/2)\%$ percentile of a standard normal distribution, i.e., $P(Z > z_{\alpha/2}) = \alpha/2$.

- z_α's are usually refered to as z critical values.

Constructing a CI

- The previous examples show the general procedure of constructing confidence intervals. Suppose $X_1, X_2, \ldots, X_n$ are the sample on which the CI for a parameter θ is to be based. Then we construct a so-called "pivotal" quantity whose distribution does not depend on parameters.

- In other words, the pivotal quantity is a function of both samples and parameters, i.e., $h(X_1, X_2, \ldots, X_n, \theta)$, and the distribution of $h(\cdot)$ does not depend on θ or any other unknowns.

- Then one can find a and b to satisfy $\mathrm{P}(a < h(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n; \theta) < b) = 1 - \alpha$, by the pivotal property, a and b do not depend on θ. Then the inequality can be manipulated to isolate θ, giving the equivalent probability statement

$$\mathrm{P}(l(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n) < \theta < u(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n)) = 1 - \alpha$$

Large-Sample Confidence Intervals for the Mean and Proportion of a *General* Population

- However, in most cases, it is impossible to locate a pivotal quantity. In the previous setting, we can do this because the unlikely assumption of knowing σ.
- We often need to resort to large-sample theory, namely Central Limit Theorem to construct CIs.
- The most common application is to construct CIs for a Population Mean and Proportion.

Key Results

- If $X_1, X_2, \ldots, X_n$ IID from a general distribution with mean μ and variance σ^2, then CLT tells us

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

or

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Key Results

- If $X_1, X_2, \ldots, X_n$ IID from a general distribution with mean μ and variance σ^2, then CLT tells us

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

or

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- Further, if we substitute σ with its estimator $\hat{\sigma}$, this still holds

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0, 1)$$

General Results

- Proposition:

  A 100(1-α)% confidence interval for the mean μ of any population when the value of σ is unknown and sample size n is sufficiently large is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right)$$

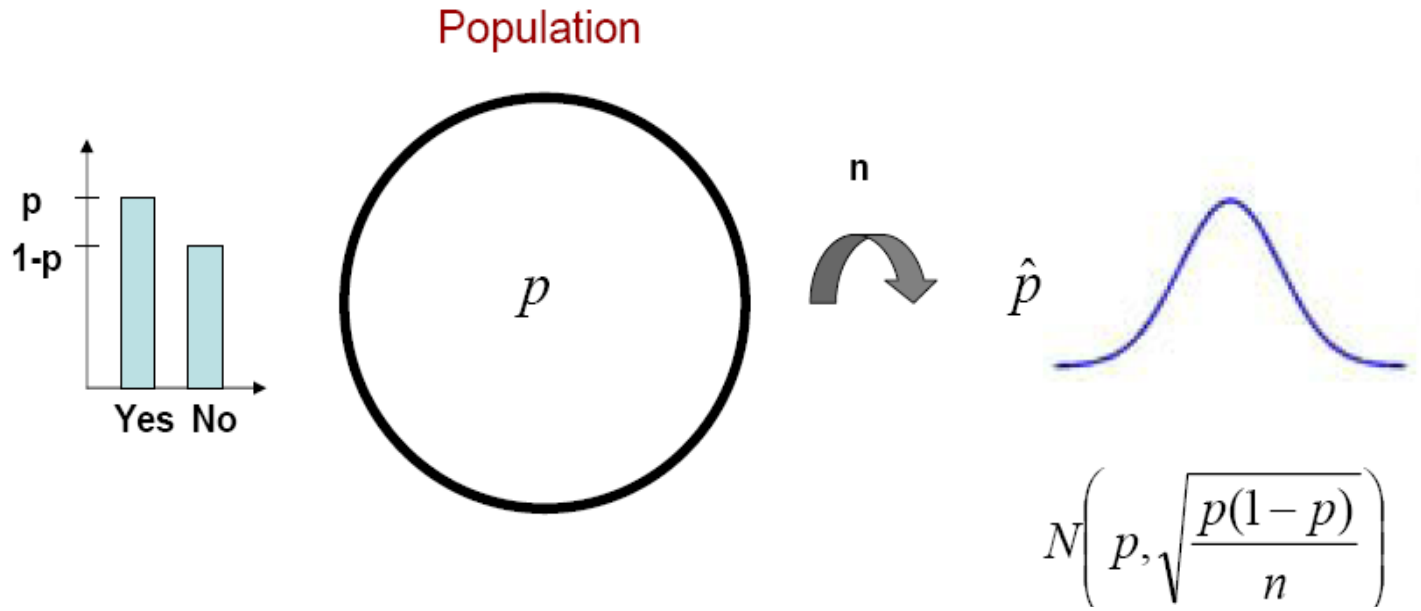
- Rule of Thumb: generally speaking, n>40 will be sufficient to justify the use of this interval. This is somewhat more conservative than the rule of thumb for the CLT, because of the additional randomness coming from $\hat{\sigma}$.

- One can also derive a similar sample size calculation formula in this case

$$n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\hat{\sigma}}{w}\right)^2$$

Proportions

- A special case of non-normal population is Bernoulli population. And the parameter of interest is the population proportion *p*.



Large Sample CI

- One can directly apply the proposition from the large sample case to construct the CI for the population proportion *p*.

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right)$$

- In this case $\bar{x} = \hat{p}$, $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$.

- If we set *q*=1-*p*, then the large sample confidence interval for *p* should be

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

- To calculate sample size: $n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sqrt{\hat{p}\hat{q}}}{w}\right)^2$

Another way

- The large sample confidence interval works fine if we have enough data. But for finite samples we can construct a better CI.

- Since in this case, we only have 1 parameter *p*, by CLT, we have

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

- If we solve the resulting quadratic function, we'll have a new confidence interval for *p*.

$$\left(\frac{\hat{p}+\frac{z_{\alpha/2}^2}{2n}-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+z_{\alpha/2}^2/n}, \frac{\hat{p}+\frac{z_{\alpha/2}^2}{2n}+z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+z_{\alpha/2}^2/n}\right)$$

Remarks

- The latter confidence interval looks complicated, but it "can be recommended for use with nearly all sample sizes and parameter values". Therefore we don't have to check for large sample conditions.

- In the latter case, we can also derive a new sample size calculation formula

$$n = \frac{2z_{\alpha/2}^2\hat{p}\hat{q} - z_{\alpha/2}^2 w^2 \pm \sqrt{4z_{\alpha/2}^4\hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2 z_{\alpha/2}^4}}{w^2}$$

  "+" sign is used!

- When sample size is large, the confidence interval we just constructed and the sample size calculation formula will be equivalent to

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right) \quad \text{and} \quad n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sqrt{\hat{p}\hat{q}}}{w}\right)^2$$

One-sided CI

- In some situations, an investigator will want only one upper bound or one lower bound for the parameter.

- Follow a similar argument as in the two-sided case, we have the following result

A large sample 100(1-α)% confidence upper bound for the mean μ is

$$\mu < \bar{x} + z_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

and a lower bound is

$$\mu > \bar{x} - z_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

A one-sided confidence bound for p results from replacing $z_{\alpha/2}$ by z_α.

CIs Based on the t Distribution

- The above discussions are based on the large-sample assumptions. But what can we do if we don't have a large sample?
- When the distribution under discussion is normal, we do have a solution, that is based on the so-called t distribution.
- Our assumption right now is $X_1, X_2, \ldots, X_n$ IID from *normal* distribution with unknown mean μ and unknown σ.

The t Distribution

- When $\bar{X}$ is the sample mean of a simple random sample from normal under the previous assumptions, then RV

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

has a probability distribution called a t distribution with $n - 1$ degrees of freedom (df). We write

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

The t Distribution

- When $\bar{X}$ is the sample mean of a simple random sample from normal under the previous assumptions, then RV

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

has a probability distribution called a t distribution with $n - 1$ degrees of freedom (df). We write

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

- The property of the t distribution
  - Bell-shaped curve centered at 0.
  - More spread-out than standard normal curve (heavy-tail).
  - When the degrees of freedom approach infinity, t distribution converges to standard normal.

t distribution table

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

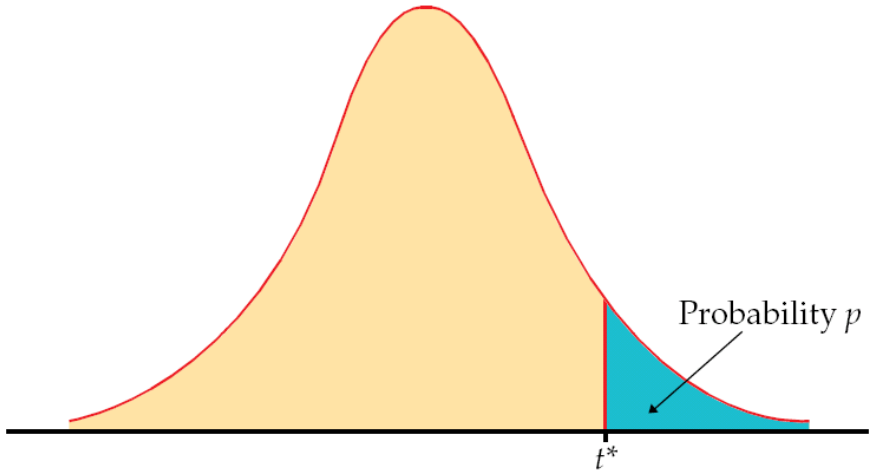


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408

Confidence Interval for μ

- Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from a simple random sample from a normal population with mean μ, then a $100(1-\alpha)\%$ confidence interval for μ is

$$(\bar{x} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}})$$

- An upper confidence interval is

$$\bar{x} + t_{\alpha,n-1}\frac{\hat{\sigma}}{\sqrt{n}}$$