Mathematical statistics

April 17^{th}, 2019

Lecture 21: Intervals based on normal distributions

Overview

8.1 Basic properties of confidence intervals (CIs)
- Interpreting CIs
- General principles to derive CI

8.2 Large-sample confidence intervals for a population mean
- Using the Central Limit Theorem to derive CIs

8.3 Intervals based on normal distribution
- Using Student's t-distribution

8.4 CIs for standard deviation

Principles for deriving CIs

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution $f(x, \theta)$, then

- Find a random variable $Y = h(X_1, X_2, \ldots, X_n; \theta)$ such that the probability distribution of Y does not depend on θ or on any other unknown parameters.
- Find constants a, b such that

$$P\left[a < h(X_1, X_2, \ldots, X_n; \theta) < b\right] = 0.95$$

- Manipulate these inequalities to isolate θ

$$P\left[\ell(X_1, X_2, \ldots, X_n) < \theta < u(X_1, X_2, \ldots, X_n)\right] = 0.95$$

Assumptions

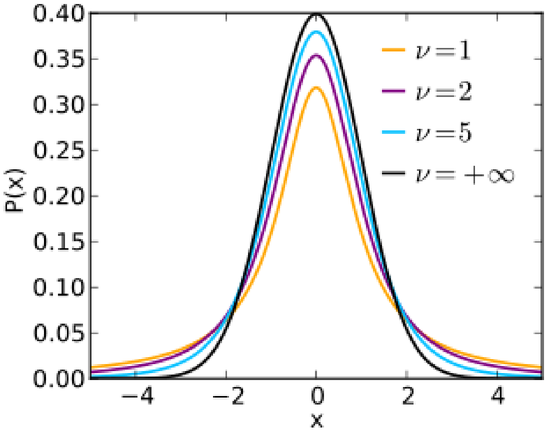
- Section 8.1
  - Normal distribution
  - σ is known
- Section 8.2
  - ~~Normal distribution~~
    $\rightarrow$ use Central Limit Theorem $\rightarrow$ needs $n > 30$
  - ~~σ is known~~
    $\rightarrow$ replace σ by s $\rightarrow$ needs $n > 40$
- Section 8.3
  - Normal distribution
  - ~~σ is known~~

$\rightarrow$ Introducing t-distribution

t distributions with degree of freedom ν

Probability density function

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



t distributions

PROPERTIES OF T DISTRIBUTIONS

1. Each t_ν curve is bell-shaped and centered at 0.
2. Each t_ν curve is more spread out than the standard normal (z) curve.
3. As ν increases, the spread of the t_ν curve decreases.
4. As $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve (so the z curve is often called the t curve with df $= \infty$).

t distributions

When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the t distribution with $n - 1$ degree of freedom (df).

Let $t_{\alpha,\nu}$ = the number on the measurement axis for which the area under the t curve with ν df to the right of $t_{\alpha,\nu}$, is α; $t_{\alpha,\nu}$ is called a **t critical value.**

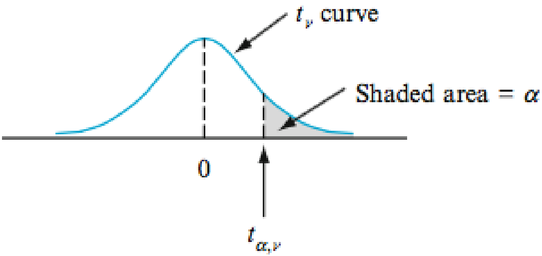


Figure 8.7 A pictorial definition of $t_{\alpha,\nu}$

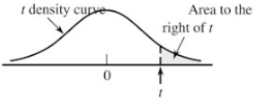
How to do computation with t distributions

- Instead of looking up the normal Z-table A3, look up the two t-tables A5 and A7.
- Idea

$$P[T \geq t_{\alpha,\nu}] = \alpha$$

- $\{$From t, find $\alpha\}$ $\rightarrow$ using table A7
- $\{$From α, find $t\}$ $\rightarrow$ using table A5

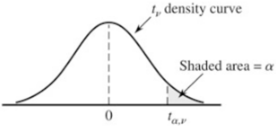
Table A.7 *t* Curve Tail Areas



t \ ν	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462.	.462	.462	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422	.422	.422	.422	.422	.422	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385	.384	.384	.384	.384	.384	.384
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348	.348	.347	.347	.347	.347	.347
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313	.313	.312	.312	.312	.312	.312
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280	.279	.279	.279	.278	.278	.278
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249	.248	.247	.247	.247	.247	.246
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220	.219	.218	.218	.218	.217	.217
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193	.192	.191	.191	.191	.190	.190
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169	.168	.167	.167	.166	.166	.165
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146	.146	.144	.144	.144	.143	.143
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127	.126	.124	.124	.124	.123	.123
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109	.108	.107	.107	.106	.105	.105
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093	.092	.091	.091	.090	.090	.089
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080	.079	.077	.077	.077	.076	.075
1.6	.178	.125	.104	.092	.085	.080	.077	.074	.072	.070	.069	.068	.067	.065	.065	.065	.064	.064
1.7	.169	.116	.094	.082	.075	.070	.065	.064	.062	.060	.059	.057	.056	.055	.055	.054	.054	.053
1.8	.161	.107	.085	.073	.066	.061	.057	.055	.053	.051	.050	.049	.048	.046	.046	.045	.045	.044
1.9	.154	.099	.077	.065	.058	.053	.050	.047	.045	.043	.042	.041	.040	.038	.038	.038	.037	.037

$\alpha \to t$

Table A.5 Critical Values for t Distributions



	α						
ν	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965

Let $\overline{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ. Then a **$100(1-\alpha)\%$ confidence interval for μ**, the **one-sample t CI**, is

$$\left(\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \tag{8.15}$$

or, more compactly, $\overline{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$.

An **upper confidence bound for μ** is

$$\overline{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$

and replacing + by − in this latter expression gives a **lower confidence bound for μ**; both have confidence level $100(1-\alpha)\%$.

Prediction intervals

Principles for deriving CIs

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution $f(x, \theta)$, then

- Find a random variable $Y = h(X_1, X_2, \ldots, X_n; \theta)$ such that the probability distribution of Y does not depend on θ or on any other unknown parameters.
- Find constants a, b such that

$$P\left[a < h(X_1, X_2, \ldots, X_n; \theta) < b\right] = 0.95$$

- Manipulate these inequalities to isolate θ

$$P\left[\ell(X_1, X_2, \ldots, X_n) < \theta < u(X_1, X_2, \ldots, X_n)\right] = 0.95$$

Settings

- We have available a random sample $X_1, X_2, \ldots, X_n$ from a normal population distribution
- We wish to predict the value of X_{n+1}, a single future observation.

This is a much more difficult problem than the problem of estimating μ

- When $n \to \infty$, $\bar{X} \to \mu$
- Even when we know μ, X_{n+1} is still random

Settings

A natural estimate of X_{n+1} is

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Question: What is the uncertainty of this estimate?

Problem

Let $X_1, X_2, \ldots, X_n$ be a sample from a normal population distribution $\mathcal{N}(\mu, \sigma)$ and X_{n+1} be an independent sample from the same distribution.

- Compute $E[\bar{X} - X_{n+1}]$ in terms of μ, σ, n
- Compute $Var[\bar{X} - X_{n+1}]$ in terms of μ, σ, n
- What is the distribution of $\bar{X} - X_{n+1}$?

Principle

If σ is known

$$\frac{\bar{X} - X_{n+1}}{\sigma\sqrt{1 + \frac{1}{n}}}$$

follows the standard normal distribution $\mathcal{N}(0, 1)$.

Principle

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1+\frac{1}{n}}} \sim t \text{ distribution with } n-1 \text{ df}$$

Prediction intervals

A prediction interval (PI) for a single observation to be selected from a normal population distribution is

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}} \qquad (8.16)$$

The *prediction level* is $100(1 - \alpha)\%$.

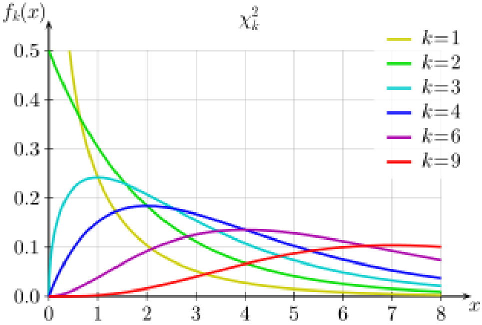
Section 6.4: Distributions based on a normal random sample

- The Chi-squared distribution
- The t distribution

Chi-squared distribution

The pdf of a Chi-squared distribution with degree of freedom ν, denoted by χ^2_ν, is

$$f(x) = \begin{cases} \dfrac{1}{2^{1/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$



Why is Chi-squared useful?

Proposition

If Z has standard normal distribution $\mathcal{Z}(0,1)$ and $X = Z^2$, then X has Chi-squared distribution with 1 degree of freedom, i.e. $X \sim \chi^2_1$ distribution.

Proposition

If $X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$ and they are independent, then

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2}$$

Why is Chi-squared useful?

Proposition

If $Z_1, Z_2, \ldots, Z_n$ are independent and each has the standard normal distribution, then

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi_n^2$$

Why is Chi-squared useful?

If $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Note that

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

- What is the distribution of the LHS?
- What is the distribution of the second term on the RHS?
- What is the distribution of

$$(n-1)\frac{S^2}{\sigma^2}$$

Why is Chi-squared useful?

If $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, then

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

t distributions

Let Z be a standard normal rv and let X be a χ^2_ν rv independent of Z. Then the t distribution with degrees of freedom ν is defined to be the distribution of the ratio

$$T = \frac{Z}{\sqrt{X/\nu}}$$

t distributions

When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the t distribution with $n-1$ degree of freedom (df).
Hint:

$$T = \frac{Z}{\sqrt{X/\nu}} \qquad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\sqrt{(n-1)\frac{S^2}{\sigma^2}/(n-1)}}.$$

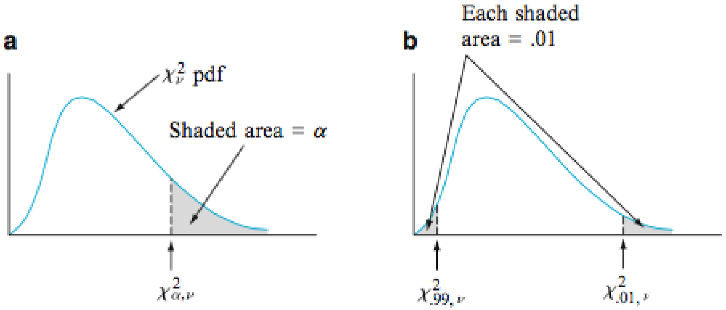
CIs for variance and standard deviation

Why is Chi-squared useful?

If $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, then

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Important: Chi-squared distribution are not symmetric



CIs for standard deviation

We have

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

Play around with these inequalities:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

CIs for standard deviation

A $100(1-\alpha)\%$ confidence interval for the variance σ^2 of a normal population has lower limit

$$(n-1)s^2/\chi^2_{\alpha/2,n-1}$$

and upper limit

$$(n-1)s^2/\chi^2_{1-\alpha/2,n-1}$$

A confidence interval for σ has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2.