Mathematical statistics

October 16^{th}, 2018

Lecture 15: Prediction intervals

Countdown to mid-term exam: 9 days

Week 1	Probability reviews
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	**Chapter 8: Confidence Intervals**
Week 10	Chapter 9: Test of Hypothesis
Week 14	Regression

Advertisement: AWM Grad School Boot Camp

- Wednesday, October 17th from 3:00-6:00 PM
- EWING 336
- Small workshops
  - how to write your personal essay
  - what to expect from the GRE
- Panel with faculties

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

Confidence Intervals

Framework

- Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution $f(x, \theta)$
- In Chapter 7, we learnt methods to construct an estimate $\hat{\theta}$ of θ
- Goal: we want to indicate the degree of uncertainty associated with this random prediction
- One way to do so is to construct a *confidence interval* $[\hat{\theta} - a, \hat{\theta} + b]$ such that

$$P[\theta \in [\hat{\theta} - a, \hat{\theta} + b]] = 95\%$$

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

95% confidence interval of the mean

- Assumptions:
  - Normal distribution
  - σ is known
- 95% confidence interval
  If after observing $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$, we compute the observed sample mean $\bar{x}$. Then

  $$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

  is a 95% confidence interval of μ

$100(1-\alpha)\%$ confidence interval

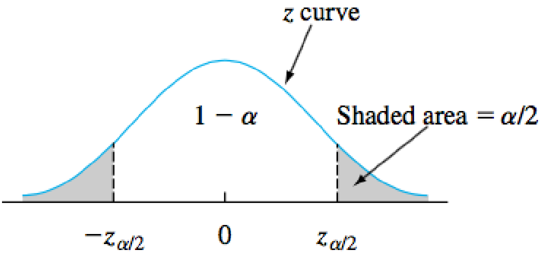


Figure 8.4 $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$

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

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{8.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

Large-sample CIs of the population mean

- Central Limit Theorem
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
is approximately normal when $n > 30$
- Moreover, when n is sufficiently large $s \approx \sigma$
- Conclusion:
$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$
is approximately normal when n is sufficiently large

If $n > 40$, we can ignore the normal assumption and replace σ by s

95% confidence interval

If after observing $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$ ($n > 40$), we compute the observed sample mean $\bar{x}$ and sample standard deviation s. Then

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$$

is a 95% confidence interval of μ

$100(1-\alpha)\%$ confidence interval

If after observing $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$ $(n > 40)$, we compute the observed sample mean $\bar{x}$ and sample standard deviation s. Then

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

is a 95% confidence interval of μ

- the population of interest is normal (i.e., $X_1, \ldots, X_n$ constitutes a random sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$).
- σ is unknown

$\rightarrow$ we want to consider cases when n is small.

t distributions

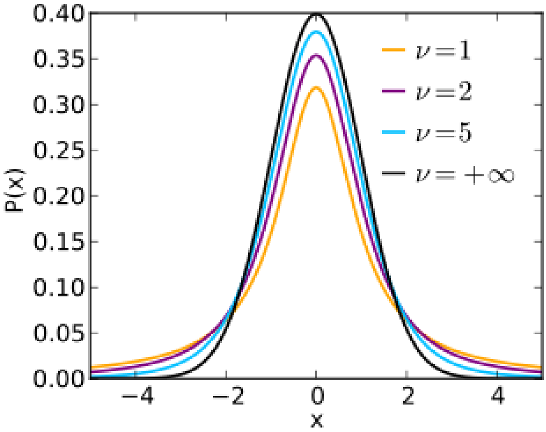
When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv

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

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

Probability density function

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



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.**

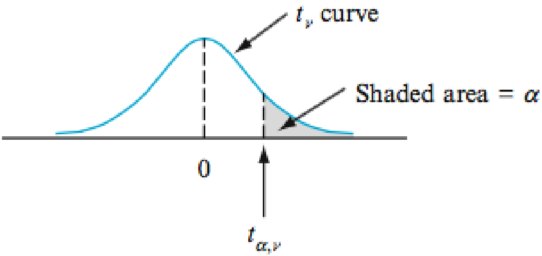


Figure 8.7 A pictorial definition of $t_{\alpha,\nu}$

Let $\bar{x}$ 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

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

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

An **upper confidence bound for μ** is

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

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

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

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

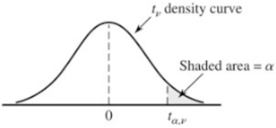
Practice problem

31. Determine the t critical value for a two-sided confidence interval in each of the following situations:

a. Confidence level = 95%, df = 10

b. Confidence level = 95%, df = 15

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

Problem

Here are the lengths (in minutes) of the 63 nine-inning games from the first week of the 2001 major league baseball season:

194	160	176	203	187	163	162	183	152	177
177	151	173	188	179	194	149	165	186	187
187	177	187	186	187	173	136	150	173	173
136	153	152	149	152	180	186	166	174	176
198	193	218	173	144	148	174	163	184	155
151	172	216	149	207	212	216	166	190	165
176	158	198							

Assume that this is a random sample of nine-inning games (the mean differs by 12 s from the mean for the whole season).

a. Give a 95% confidence interval for the population mean.

b. Give a 95% prediction interval for the length of the next nine-inning game. On the first day of the next week, Boston beat Tampa Bay 3–0 in a nine-inning game of 152 min. Is this within the prediction interval?

$\Phi(z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997

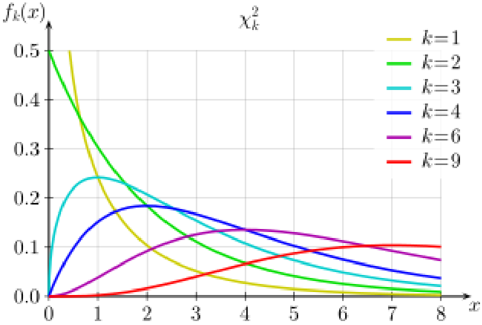
Section 6.4: Distributions based on a normal random sample

- The Chi-squared distribution
- The t distribution
- The F 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

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

F distributions

Let X_1 and X_2 be independent chi-squared random variables with ν_1 and ν_2 degrees of freedom, respectively. The F_{ν_1,ν_2} distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom is defined to be the distribution of the ratio

$$\frac{X_1/\nu_1}{X_2/\nu_2}$$

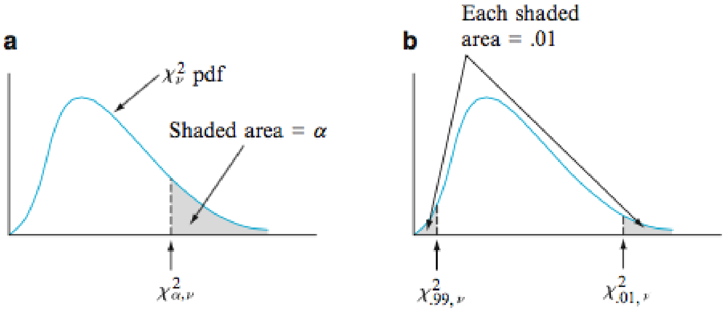
CIs for variance and standard deviation

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.