Basic Statistics and Probability

Chapter 7:
Inferences Based on a Single Sample: Estimation with Confidence Intervals

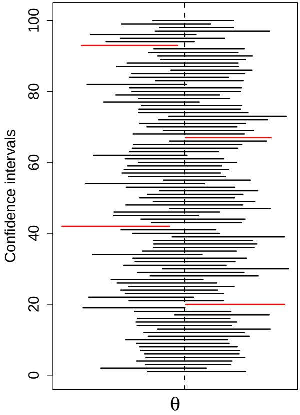
- Confidence Interval
- Confidence Interval for a Population Mean
- Confidence Interval for the Mean of a Normal Population
- Confidence Interval for a Population Proportion
- Determining the Sample Size

- In Ch. 6 we saw that a point estimator of a target population parameter θ is a specific quantity $\hat{\theta}$ estimating θ and calculated from the sample $X_1, \ldots, X_n$.
- A point estimate will "always" carry an error. We seek to measure the uncertainty inherent to the point estimator.
- An interval estimator or confidence interval for the target population parameter θ is a whole interval of values estimating θ and calculated from the sample $X_1, \ldots, X_n$.
- Interval estimation gives an interval containing the parameter θ with a predetermined high confidence level.
- The confidence level $1 - \alpha$ is a measure of our degree of certainty that θ will be in the interval.
- We will talk about, say, 90%, 95%, 99% confidence intervals (intervals with a confidence level of 90%, 95%, 99% or confidence coefficient of 0.9, 0.95, 0.99).

Interpretation of the confidence level

Suppose we observe 100 samples of size n from a r.v. X whose distribution depends on a parameter θ. Then we construct the corresponding 100 confidence intervals for θ with confidence level $1-\alpha$, $\mathrm{CI}_{1-\alpha}(\theta)$.

$$
\begin{aligned}
\text{Sample 1} &\quad\rightarrow\quad \mathrm{CI}_{1-\alpha}^{(1)}(\theta)\\
\text{Sample 2} &\quad\rightarrow\quad \mathrm{CI}_{1-\alpha}^{(2)}(\theta)\\
&\quad\vdots\\
\text{Sample 100} &\quad\rightarrow\quad \mathrm{CI}_{1-\alpha}^{(100)}(\theta)
\end{aligned}
$$



The parameter θ will be in approximately $(1-\alpha)100$ of them.

Confidence Interval for a Population Mean

Let X be a random variable with mean μ and standard deviation σ. For a sample of X with sample mean $\bar{X}$, by the CLT we know that, for a large sample size n, the approximate distribution of the z-statistic

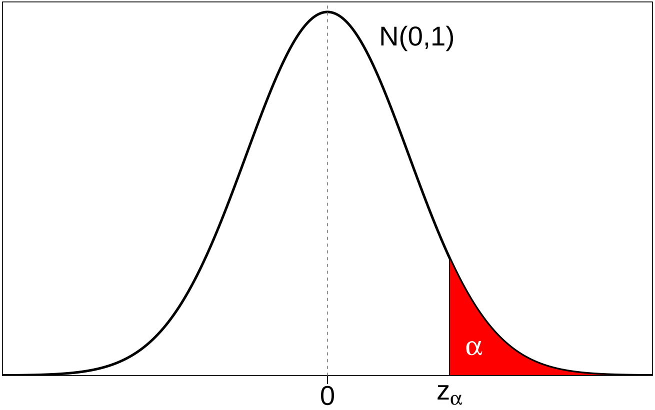
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{if } \sigma \text{ unknown}}{\simeq} \frac{\bar{X} - \mu}{s/\sqrt{n}}.$$

is $N(0,1)$.

Then, for large n ($n \geq 20$), a confidence interval for μ at the confidence level $1-\alpha$ is

$$\text{CI}_{1-\alpha}(\mu) = \left(\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

z_α is that value leaving to its right an area equal to α in the N(0,1) density:



Check on your own that:

$$z_{0.05} = 1.645 \qquad z_{0.01} = 2.33 \qquad z_{0.005} = 2.575$$

Exercise in McClave & Sincich: Latex allergy in health care workers. Health care workers who use latex gloves with glove powder may develop a latex allergy. Symptoms of a latex allergy include conjunctivitis, hand eczema, nasal congestion, a skin rash, and shortness of breath. Each in a sample of 46 hospital employees who were diagnosed with latex allergy reported on their exposure to latex gloves (*Current Allergy & Clinical Immunology*, Mar. 2004). Summary statistics for the number of latex gloves used per week are $\bar{x} = 19.3$ and $s = 11.9$.

a. Give a point estimate for the average number of latex gloves used per week by all health care workers with a latex allergy.

b. Form a 95% confidence interval for the average number of latex gloves used per week by all health care workers with a latex allergy.

Confidence Interval for the Mean of a Normal Population

If we know that the variable X follows a normal distribution with mean μ (unknown) and standard deviation σ (unknown), then it is possible to compute confidence intervals for the population mean μ for any sample size n (even small ones).

To this end, we introduce a new probability distribution, called Student's t.

The t distribution is a continuous one, with a density which is symmetric with respecto to 0, but with tails heavier than those of the standard normal.

The t distribution depends on a parameter called the degrees of freedom (d.f.). A t distribution with, say, 3 d.f. is denoted by t_3.

As the d.f. increase the tails of the t_{df} get lighter. When d.f.≥ 30, the t_{df} can be approximated by a N(0,1).

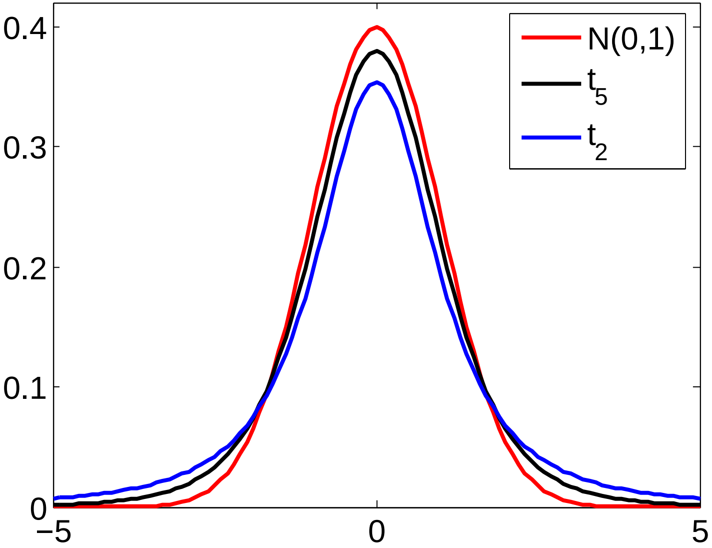
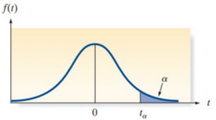
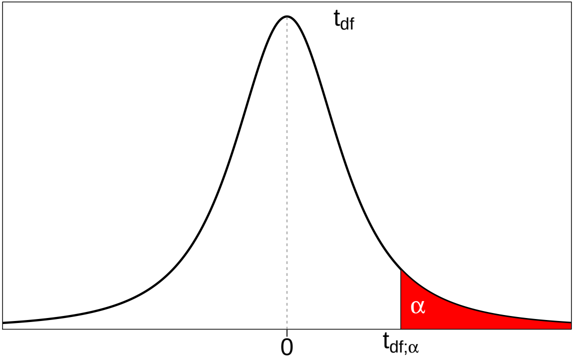

0.4
0.3
0.2
0.1
0
−5
0
5
N(0,1)
t_5
t_2

Table III Critical Values of t



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
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
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

$t_{\text{df};\alpha}$ is the value leaving to its right an area α in the t_{df} density:



If df $\geq$ 30, then $t_{\text{df};\alpha} \simeq z_\alpha$.

Check on your own that:

$$t_{10;0.05} = 1.812 \qquad t_{4;0.01} = 3.747 \qquad t_{5;0.005} = 4.032$$

If $X \sim N(\mu, \sigma)$ and $X_1, \ldots, X_n$ is a sample from X with sample mean $\bar{X}$, then the t-statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

follows a t_{n-1} distribution. This result holds for any n.

Then a confidence interval for the mean μ of a normal population at the confidence level $1 - \alpha$ is

$$\text{CI}_{1-\alpha}(\mu) = \left(\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \right)$$

Exercise in McClave & Sincich: Radon exposure in Egyptian tombs. Many ancient Egyptian tombs were cut from limestone rock that contained uranium. Since most tombs are not well-ventilated, guards, tour guides, and visitors may be exposed to deadly radon gas. In *Radiation Protection Dosimetry* (Dec. 2010), a study of radon exposure in tombs in the Valley of Kings, Luxor, Egypt (recently opened for public tours), was conducted. The radon levels – measured in becquerels per cubic meter (Bq/m3) – in the inner chambers of a sample of 12 tombs were determined. For this data, assume that $\bar{x} = 3,643$ Bq/m^3 and $s = 1,187$ Bq/m^3. Use this information to estimate, with 95% confidence, the true mean level of radon exposure in tombs in the Valley of Kings.

Confidence Interval for a Population Proportion

Let X be a Bernoulli(p) r.v.

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

For a sample of size n from X we compute its sample mean, $\hat{p} = \bar{X}$. By the CLT, for large n, the z-statistic

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/\sqrt{n}}}$$

is approximately $N(0,1)$.

Then, for large n ($n \geq 20$) and $0.1 \leq p \leq 0.9$, a confidence interval for p at the confidence level $1-\alpha$ is

$$\mathrm{CI}_{1-\alpha}(p) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right).$$

Exercise in McClave & Sincich: Paying for music downloads. If you use the Internet, have you ever paid to access or download music? This was one of the questions of interest in a recent *Pew Internet and American Life Project Survey* (Oct. 2010). Telephone interviews were conducted on a representative sample of 1,003 adults living in the United States. For this sample, 506 adults stated that they have paid to download music.

a. Use the survey information to find a point estimate for the true proportion of U.S. adults who have paid to download music.

b. Find an interval estimate for the proportion, part **a**. Use a 90% confidence interval.

c. Give a practical interpretation of the interval, part **b**. Your answer should begin with "We are 90% confident ..."

The appropriate sample size n for making an inference about a population mean or proportion depends on the desired certainty.

The reliability of a confidence interval for the population mean μ or proportion p is given by the sampling error SE within which we want to estimate μ or p with that confidence level:

$$\text{SE} = \text{Half-width of the confidence interval.}$$

Example (Latex allergy in health care workers):

Suppose we want to estimate μ, the expected number of latex gloves used per week by health care workers with a latex allergy, with a confidence level of 95% and a sampling error SE of 1. Which is the minimum required sample size?

Example (Radon exposure in Egyptian tombs):

What sample size should they use if the researchers want to estimate the mean level of radon exposure in those tombs to within 300 Bq/m^3 of its true value?

Example (Paying for music downloads):

How many telephone interviews should be conducted in order to estimate the proportion p of U.S. adults who have paid to download music to within 0.01 with 90% confidence?