Parameter Estimation

Goals of Parameter Estimation

- Let $X_1, \dots, X_n$ be a random sample from a distribution F_θ with unknown parameter vector θ. Whereas in probability theory it is usual to suppose that all of the parameters of a distribution are known, the opposite is true in statistics and data engineering, where a central problem is to use the observed data to make inferences about the unknown parameters
- One frequent application is to estimate the mean and the variance
- Two types of estimates: a) point estimate b) level estimate
- We start with point estimate

Sampling Distribution of Mean

Consider a population with population mean μ and population variance σ^2 . We wish to estimate them from n random samples of the population $X_1, X_2, \dots, X_n$. The sample mean is

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

$$E[\bar{X}] = E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \mu$$

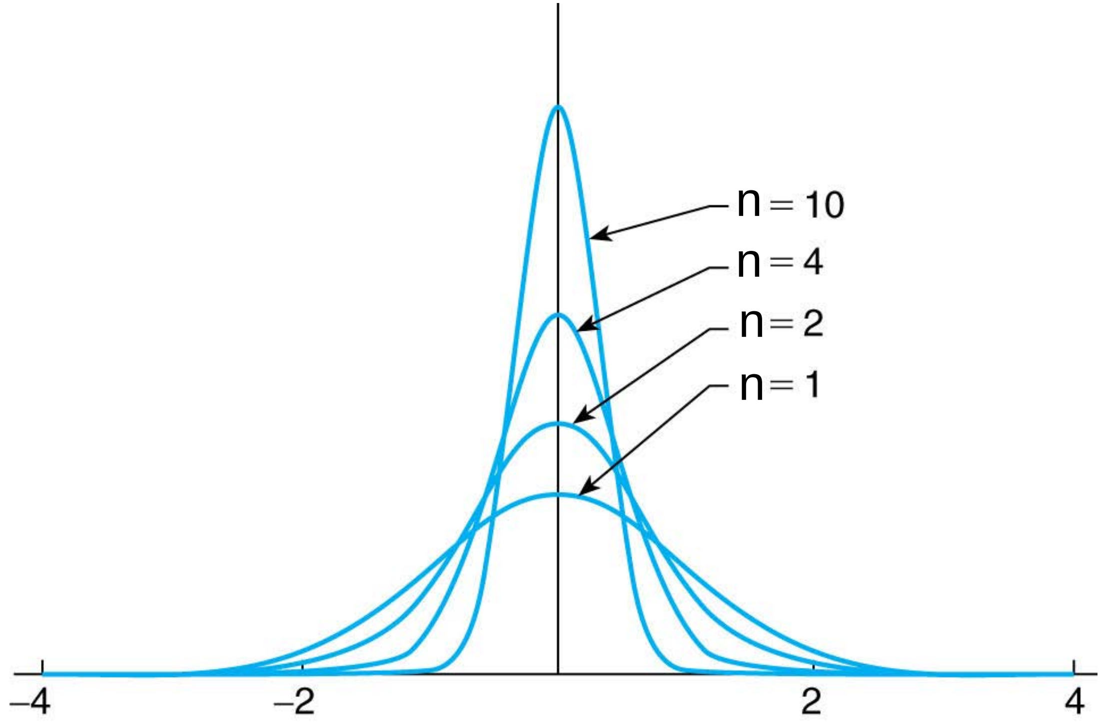
$$Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right)$$

$$= \frac{1}{n^2}[Var(X_1) + \cdots + Var(X_n)] \qquad by\ independence$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\bar{X}$ is centered about the mean μ

Its spread becomes more and more reduced as the sample size n increases



Probability density function of the sample mean from a standard normal distribution

Sample Mean is an Unbiased Estimator

The sample mean $\bar{X}$ is a point estimator of the population mean μ.

An estimator is unbiased if its expected value is always equal to the value of the parameter it is attempting to estimate. Otherwise the difference is the bias (which may be positive or negative). Since

$$E[\bar{X}] = \mu$$

The sample mean is an unbiased estimator of the true mean

Sample Variance is an Unbiased Estimator

Sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Using $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2$

$$(n-1)S^2 = \sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2$$

$$(n-1)E[S^2] = \sum_{i=1}^{n} E[{X_i}^2] - nE[\bar{X}^2] = nE\left[{X_1}^2\right] - nE[\bar{X}^2]$$

Using $Var(X) = E[X^2] - (E[X])^2$

$$nE\left[{X_1}^2\right] - nE[\bar{X}^2]$$
$$= n(Var(X_1) + (E[X_1])^2) - n(Var(\bar{X}) + (E[\bar{X}])^2)$$
$$= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2$$

Hence

$$E[S^2] = \sigma^2$$

The sample variance is an unbiased estimator of the population variance. Note that the denominator should be $(n-1)$ to make S^2 unbiased

Bayesian Inference Revisited

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{P(EF)}{P(E)}\frac{P(E)}{P(F)} = \frac{P(F|E)P(E)}{P(F)}$$

We often know $P(data|hypothesis)$. However, we would like to know $P(hypothesis|data)$. Bayesian inference enables us to do so by

Posterior Probability

Likelihood

Prior Probability

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

Evidence
(also called marginal likelihood)

Maximum Likelihood Estimation Revisited

We wish to find the hypothesis that maximize $P(hypothesis|data)$. Note that $P(data)$ does not affect the result. If we have no information about $P(hypothesis)$ (i.e., each hypothesis is equally likely), then we can find the hypothesis that maximizes the likelihood function

$P(data|hypothesis)$

If the hypothesis is a parameter, it gives the maximum likelihood estimator of the parameter.

Example

Suppose that n independent trials, each of which has a success probability of p, is performed, derive an expression for the maximum likelihood estimator of p

The data consists of the values $x_1, \ldots, x_n$ of the Bernoulli random variable $X_1, \ldots, X_n$ such that

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

This can be expressed as

$$P\{X_i = x\} = p^x(1-p)^{1-x} \qquad x = 0, 1$$

The maximum likelihood estimator of p is an expression of p which maximizes the probability

$$f(x_1, \dots, x_n | p)$$

$$f(x_1, \dots, x_n | p) = P\{X_1 = x_1, \dots, X_n = x_n | p\}$$

Since the trials are independent,

$$f(x_1, \dots, x_n | p) = p^{x_1}(1-p)^{1-x_1} \dots p^{x_n}(1-p)^{1-x_n}$$
$$= p^{\sum_1^n x_i}(1-p)^{n-\sum_1^n x_i} \qquad x_i = 0,1$$

Take log,

$$\log f(x_1, \ldots, x_n|p) = \sum_1^n x_i \log p + (n - \sum_1^n x_i) \log(1-p)$$

Since we wish to find p which maximizes $f(x_1, \ldots, x_n|p)$, we compute

$$\frac{d}{dp} \log f(x_1, \ldots, x_n|p)$$

The maximum (or minimum) is obtained by setting it to 0

$$\frac{d}{dp} \log f(x_1, \ldots, x_n|p) = \frac{\sum_1^n x_i}{p} - \frac{(n - \sum_1^n x_i)}{1-p} = 0$$

This gives

$$\frac{\sum_1^n x_i}{p} = \frac{n - \sum_1^n x_i}{1-p} \Rightarrow p = \frac{\sum_{i=1}^n x_i}{n}$$

Hence the maximum likelihood estimator of p is the sampled mean number of successes

Bayes Estimator

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

Bayes estimator computes the mean of the posterior distrbution $P(hypothesis|data)$, i.e., $E[hypothesis|data]$

This requires the probability distribution $P(hypothesis)$ to be known. This enables $P(data)$ to be calculated by integration

Example

An example is given in Example 7.8a (pg. 280) of text

The maximum likelihood method and the Bayes estimation method are well known methods for point estimation

Below we study interval estimation, which requires us to assume that the random variable belongs to a known distribution. One common distribution that is frequently assumed is the normal distribution

Normal Random Variables

- Also called Gaussian random variable
- A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim N(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

- The normal density $f(x)$ is a bell-shaped curve that is symmetric about μ and that attains its maximum value $\frac{1}{\sqrt{2\pi}\sigma} \approx \frac{0.399}{\sigma}$ at $x = \mu$

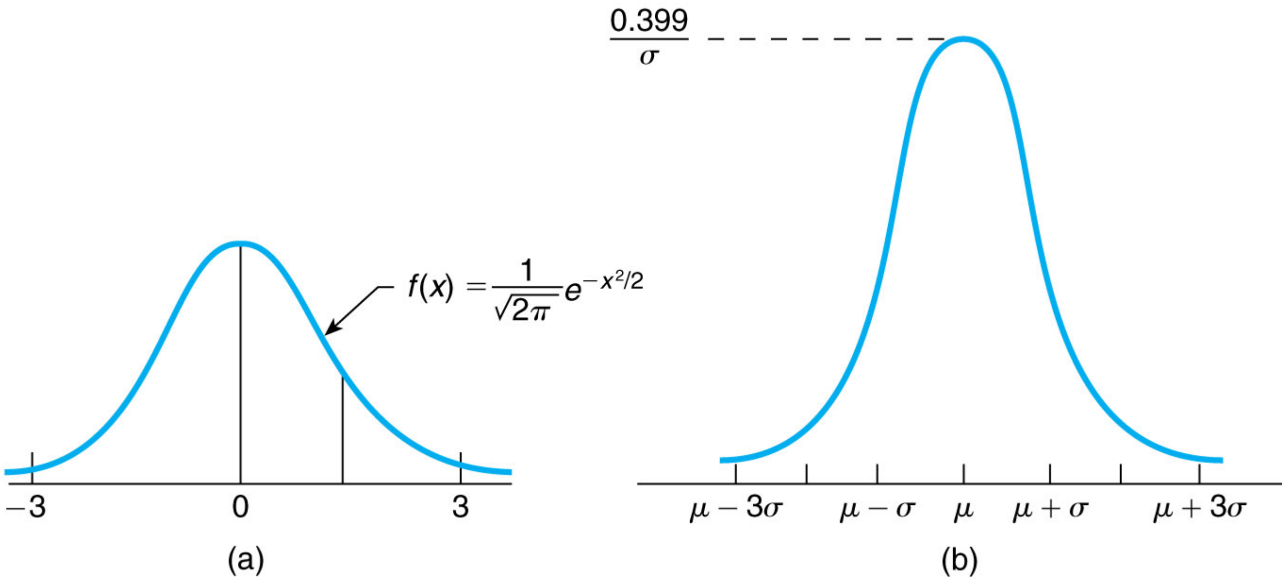


FIGURE *The normal density function (a) with $\mu = 0$, $\sigma = 1$ and (b) with arbitrary μ and σ^2.*

Standard Normal Distribution (also called Z distribution)

If $X \sim N(\mu, \sigma^2)$,

$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with mean 0 and variance 1, as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad \Rightarrow \quad f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Such a random variable Z is said to have a standard normal distribution

Normality Test

- Normality tests are used to check whether the distribution is approximately normal
- A rule of thumb is to compute the skew of the distribution. If skew > $2\sqrt{6/n}$, then the skew is significant and the distribution is not normal
- There are other normality tests such as Shapiro-Wilk test, Jarque-Bera test and Kolmogorov–Smirnov test

Central Limit Theorem

Let $X_1, X_2, \ldots X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables each having mean μ and variance σ^2. Then for large n, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$

(This is easily remembered by using linearity of the mean and the rule that the variance of the sum is the sum of the variance for independent variables)

Central Limit Theorem (alternative statement)

Let $X_1, X_2, \ldots X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables each having mean μ and variance σ^2. Then for large n, the distribution of

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

is approximately normal with mean μ and variance σ^2/n

(for variance, make use of the fact that $Var(aX + b) = a^2$Var(X))

Physical meaning

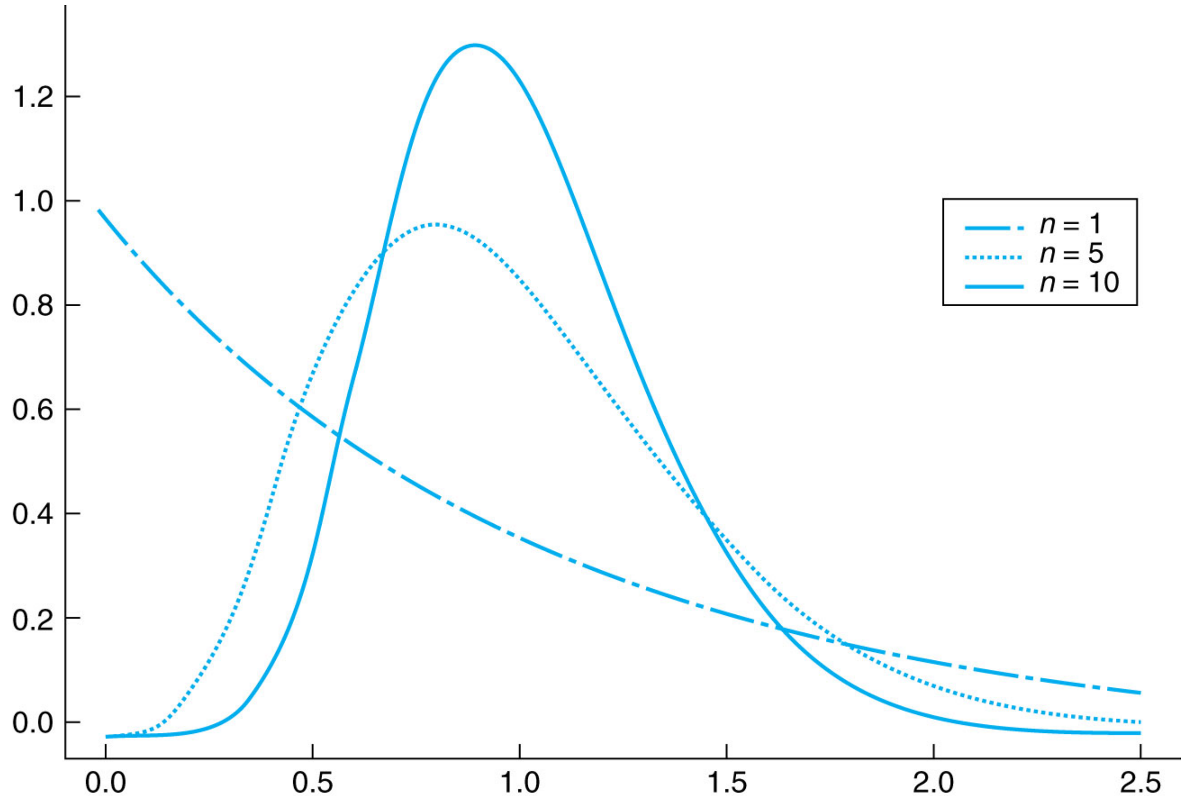
- The theorem asserts that the sum of a large number of i.i.d. random variables has a distribution that is approximately normal
- It also asserts that the sample mean of a large number of i.i.d. random variables has a distribution that is approximately normal
- In nature, approximately normal distributions are ubiquitous. The theorem provides one explanation for this phenomenon

How large a sample is needed

A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30. That is, practically specking, no matter how non-normal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal.

Of course, if the distribution is normal, then the distribution is normal for whatever sample size

An example is
http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html



Probability distribution of the average of n exponential random variables with mean 1

Example

The heights of a population of students have mean 167 cm and standard deviation 27 cm.

Assume the height of a student is i.i.d.

If a sample of 36 students is chosen, approximate the probability that the sample mean of their heights lies between 163 cm and 171 cm.

Do you think the height of a student is i.i.d.? Why? (For discussion)

Assume the height of students X is an i.i.d. random variable. The central limit theorem suggests that $\bar{X}$ is approximately normal with mean 167 cm and standard deviation $\frac{27}{\sqrt{36}} = 4.5$ cm

sd/sqr(n

Converting to standard normal variable,

$$P\{163 < \bar{X} < 171\} = P\left\{\frac{163 - 167}{4.5} < \frac{\bar{X} - 167}{4.5} < \frac{171 - 167}{4.5}\right\}$$

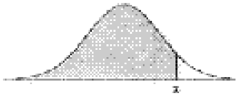
$$= P\left\{-0.8889 < \frac{\bar{X} - 167}{4.5} < 0.8889\right\}$$

$$\approx 2P\{Z < 0.89\} - 1$$

$$= 2(0.8133) - 1 = 0.6266$$

Z Distribution Table

Tables of the Normal Distribution



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Example

Show that when n is large, the binomial distribution with parameter (n, p) can be approximated by a normal distribution with mean np and variance $np(1-p)$

The binomial random variable X is the sum of n i.i.d. random Variables (i.e., Bernoulli random variables.)

$$X = \sum_{i=1}^{n} X_i$$

For a Bernoulli random variable X_i,

$\mu = E[X] = p$ $\quad\quad\quad\quad$ $\sigma^2 = Var[X_i] = p(1-p)$

By the central limit theorem (first version), when n is large, it can be approximated by a normal random variable with mean $n\mu = np$ and variance $n\sigma^2 = np(1-p)$

The approximation speeds up calculations and also make analytical calculations in theory easier

(Note that when n is large and p is small, the binomial distribution may be approximated by the Poisson distribution)

Interval Estimate

- Specify an interval for which we have a certain degree of confidence that a parameter, e.g. population mean μ, lies
- We wish to obtain "a $100(1-\alpha)\%$ confidence interval" of a parameter. For example, commonly $\alpha = 0.05$, and we wish to obtain a 95% confidence interval at which μ lies using the sampled data

Example: Method to obtain a 95% confidence interval of μ for a normal distribution, assuming value of the variance σ^2 is known

The point estimator $\bar{X}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$$

has a standard normal distribution. Therefore

$$P\left\{-1.96 < \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} < 1.96\right\} = 0.95$$

Simplifying,

$$P\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

We now observe the sample and it turns out that $\bar{X} = \bar{x}$, then we say that "with 95% confidence"

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

or

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

Statistics Uses and Misuses

- A common misconception is that the true parameter has 95% probability of lying in the confidence interval. This is wrong
- The reason is that the confidence interval is computed from the statistical data $x_1, \ldots, x_n$. No random variable is involved and hence no probability can be computed
- The true physical meaning of the $(1-\alpha)\%$ confidence interval is that if we repeat the procedure of sampling many times, in $(1-\alpha)\%$ times the true parameter will lie within the confidence interval

Chi-Square Distribution (χ^2 Distribution)

i.i.d.

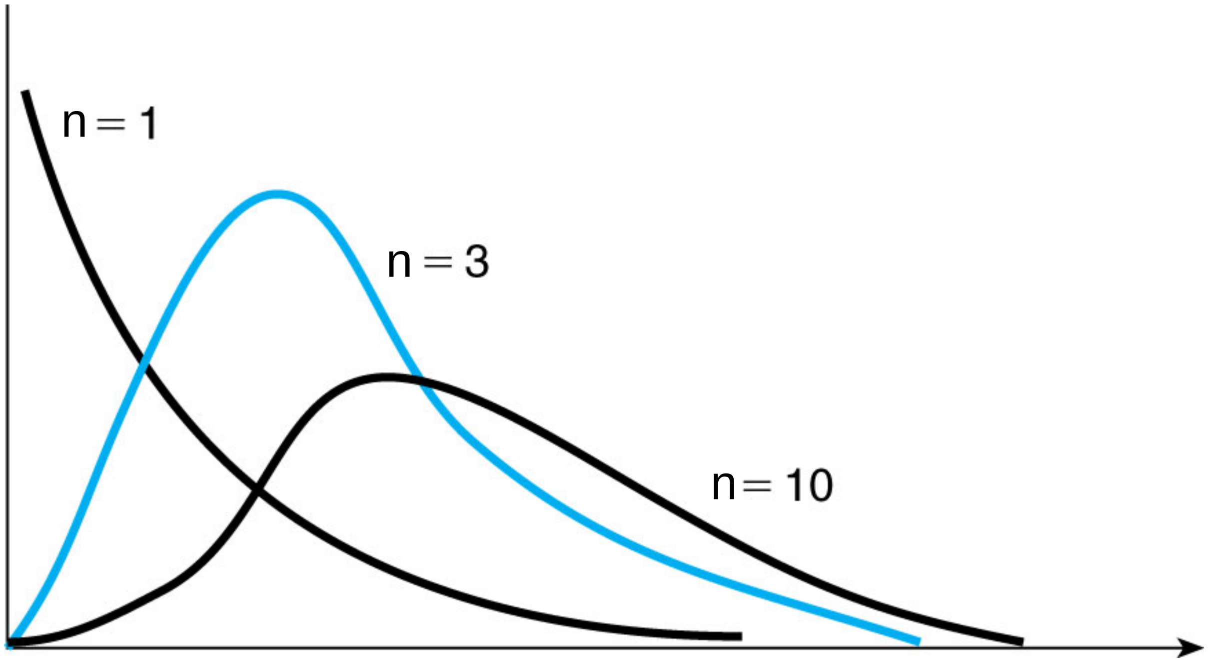
If $Z_1, Z_2, \dots, Z_n$ are independent standard normal random variables, then X, defined by

$$X = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_n}^2$$

has a chi-square distribution with n degree of freedom. We use the notation

$$X \sim {\chi_n}^2$$

E[Z^2] = var(Z)



Chi-square density function with n degrees of freedom

t- Distribution (Student's t-distribution)

If Z and ${\chi_n}^2$ are <u>independent</u> random variables, with Z having a standard normal distribution and ${\chi_n}^2$ having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{{\chi_n}^2/n}}$$

is said to have a t-distribution with n degrees of freedom

Like the normal distribution, the t-distribution is symmetric about the mean, and the mean is 0.

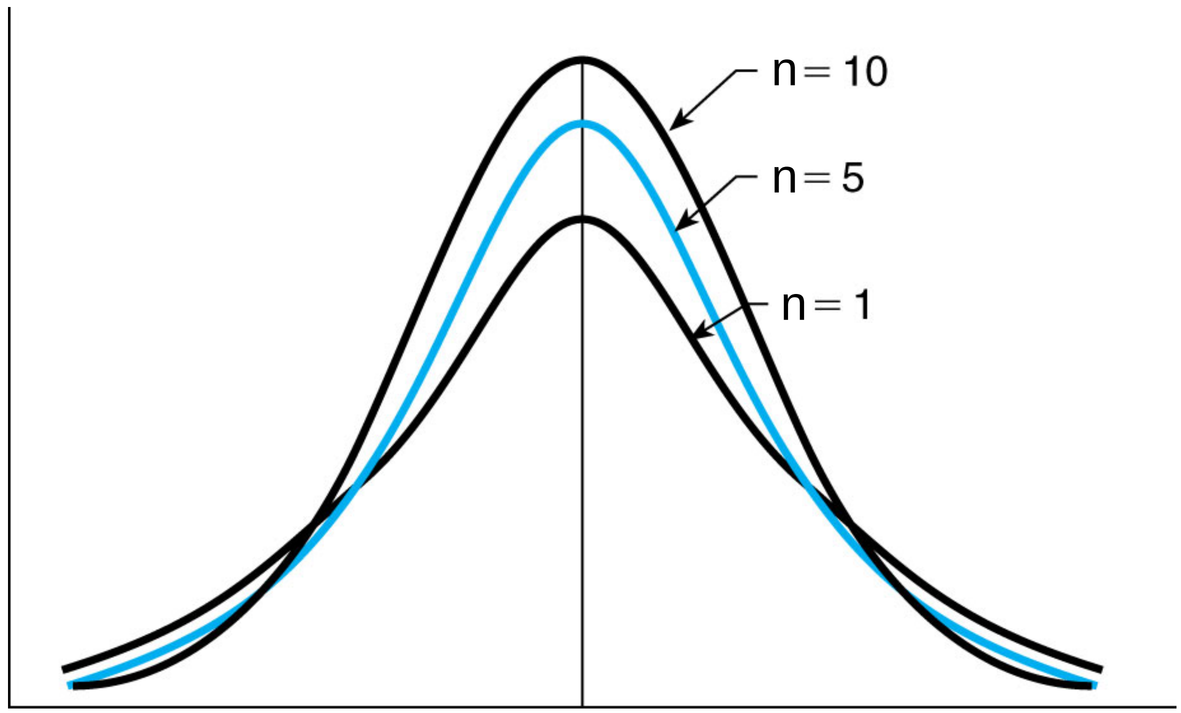
As n becomes larger, it becomes more and more like a standard normal distribution. Note that

$$\frac{{\chi_n}^2}{n} = \frac{{Z_1}^2 + \cdots + {Z_n}^2}{n}$$

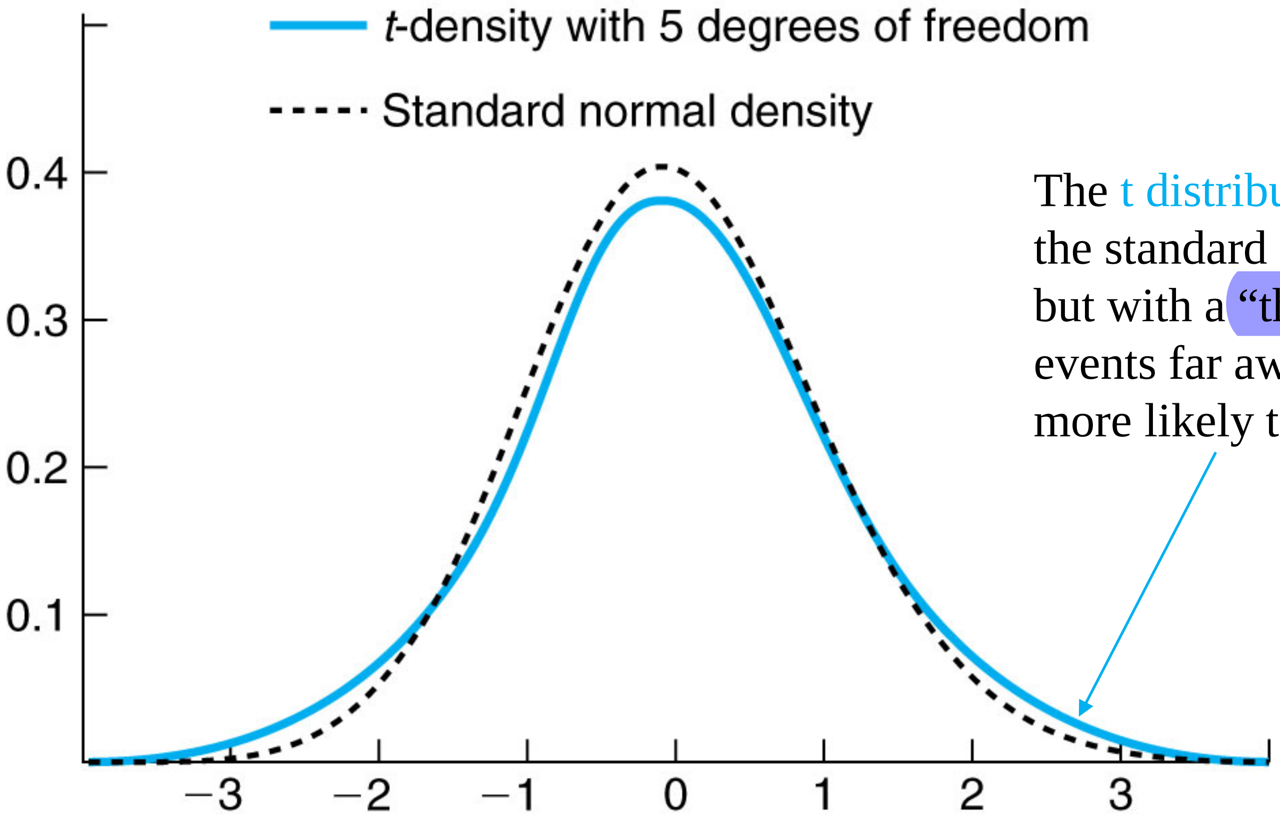
where $Z_1, \ldots, Z_n$ are independent standard normal random variables. By the weak law of large numbers, for large n, $\frac{{\chi_n}^2}{n}$ will with probability close to 1, be approximately equal to $E\left[{Z_i}^2\right] = 1$. Hence for large n,

$$T_n = \frac{Z}{\sqrt{{\chi_n}^2/n}} \approx Z$$

As n becomes large, the t distribution looks more and more like a standard normal distribution Z



Probability density function Tn



The t distribution has a shape like the standard normal distribution Z but with a "thicker tail". It means that events far away from the mean is more likely to occur

The case of unknown variance

Suppose $X_1, \ldots, X_n$ is a sample from a normal distribution with unknown mean μ and unknown variance σ^2. The sample variance is

$$S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(\text{n-1})$$

It can be shown that (Theorem 6.5.1 and Corollary 6.5.2 of text)

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim T_{n-1}$$

i.e., $\sqrt{n}\frac{(\bar{X}-\mu)}{S}$ has a t-distribution with $n-1$ degrees of freedom

Confidence Interval for the Mean of a Normal Distribution

From the symmetry of the t-density function, for $\alpha \in (0,0.5)$

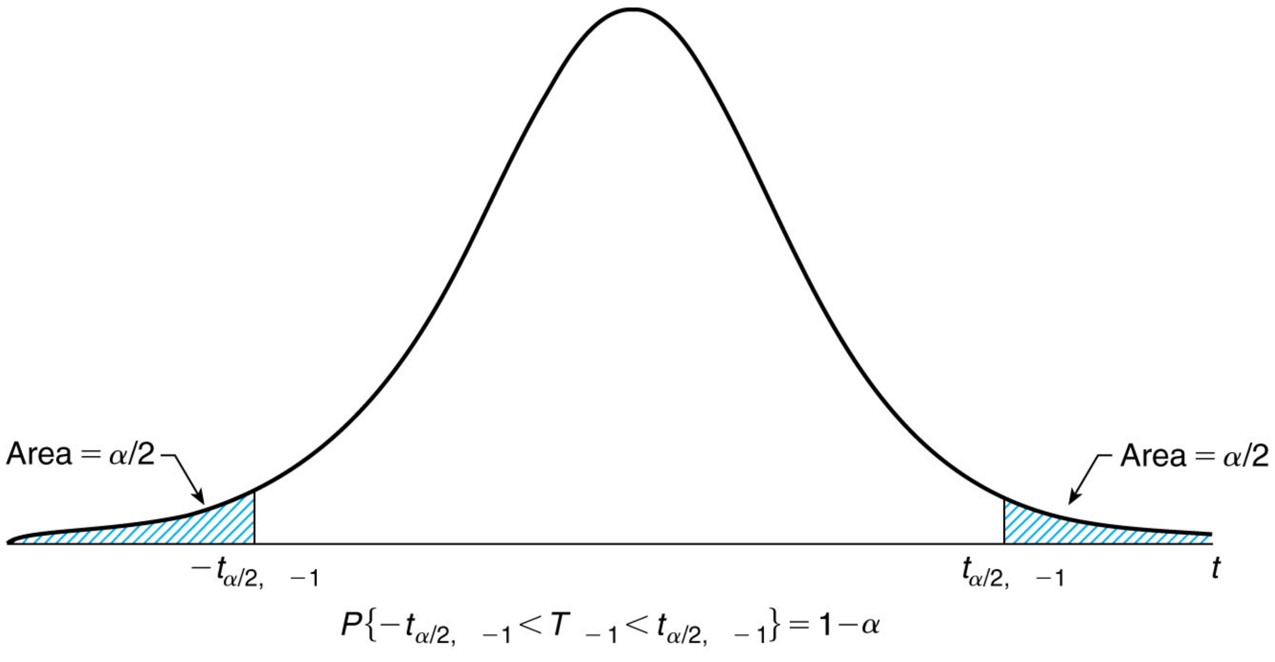
$$P\left\{-t_{\frac{\alpha}{2},n-1} < \sqrt{n}\frac{(\bar{X}-\mu)}{S} < t_{\frac{\alpha}{2},n-1}\right\} = 1-\alpha$$

or

$$P\left\{\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

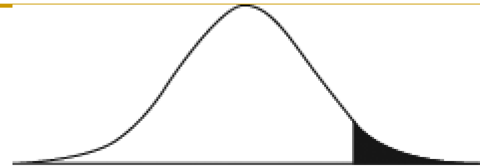
We now observe that $\bar{X} = \bar{x}$ and $S = s$, then "with $100(1-\alpha)$ % confidence"

$$\mu \in \left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)$$



Confidence interval using t
probability density function

t- distribution table



v is the degree of freedom

	Tail probability									
ν	**0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Example

In a election poll of 10 voters, the sample mean of a candidate is 20% and the sample variance is 4. Estimate the 95% confidence interval of the mean. What assumption have you made?

$n = 10$. degree of freedom (d.f.) = 9 $s = 2$ $\alpha = 0.05$

sample -> 10-1

$t_{.025,9} = 2.262$

(1-.95)/2

$$\mu \in \left(\bar{x} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}} \right)$$

$$= \left(20 - 2.262 \frac{2}{\sqrt{10}}, 20 + 2.262 \frac{2}{\sqrt{10}} \right) \approx 20 \pm 1.43\%$$

One has implicitly assume that the distribution is normal!

In the previous example, if the number of voters increases to 1000, what is the corresponding confidence interval?

$n = 1000$, d.f. = 999. The t-distribution table does not have this value and we use d.f. = ∞. $t_{0.025,\infty} = 1.96$

$$\mu \in \left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)$$
$$= \left(20 - 1.96\frac{2}{\sqrt{1000}}, 20 + 1.96\frac{2}{\sqrt{1000}}\right) = 20 \pm 0.124\%$$

Note that if we assume s^2 is the true variance σ^2 and use the Z distribution,

1. unknown variance
2. the random variable is i.i.d.
3*. each of the random variable is normal

$$\mu \in \left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$
$$= \left(20 - 1.96\frac{2}{\sqrt{1000}}, 20 + 1.96\frac{2}{\sqrt{1000}}\right) = 20 \pm 0.124\%$$

Thus for $d.f. \gg 130$, the t-distribution can be approximated by the Z distribution. This is why when the sample size is large, we can estimate assuming that the sampling variance is the true variance and then use the Z distribution

Note also that the Z distribution gives you a tighter confidence interval

References

- Text Ch. 5-7
- Normality Test
  - https://en.wikipedia.org/wiki/Normality_test
  - https://www.youtube.com/watch?v=8EXZrb9TrZg