Non-normal and Unknown Variance

- Previously we constructed a confidence interval for normal population mean with known variance. The next question would then be, what if we don't have normality and what if we don't know the underlying variance?

- If we have large enough sample size, the celebrated CLT can help us construct a confidence interval for the mean parameter of a population with unknown distribution and unknown variance. Consider the following quantity

$$\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} = \underbrace{\boxed{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}}_{\xrightarrow{\text{CLT}}\ \boxed{N(0,1)}} \cdot \underbrace{\boxed{\frac{\sigma}{\hat{\sigma}}}}_{\xrightarrow{\text{LLN}}\ \boxed{1}}$$

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

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

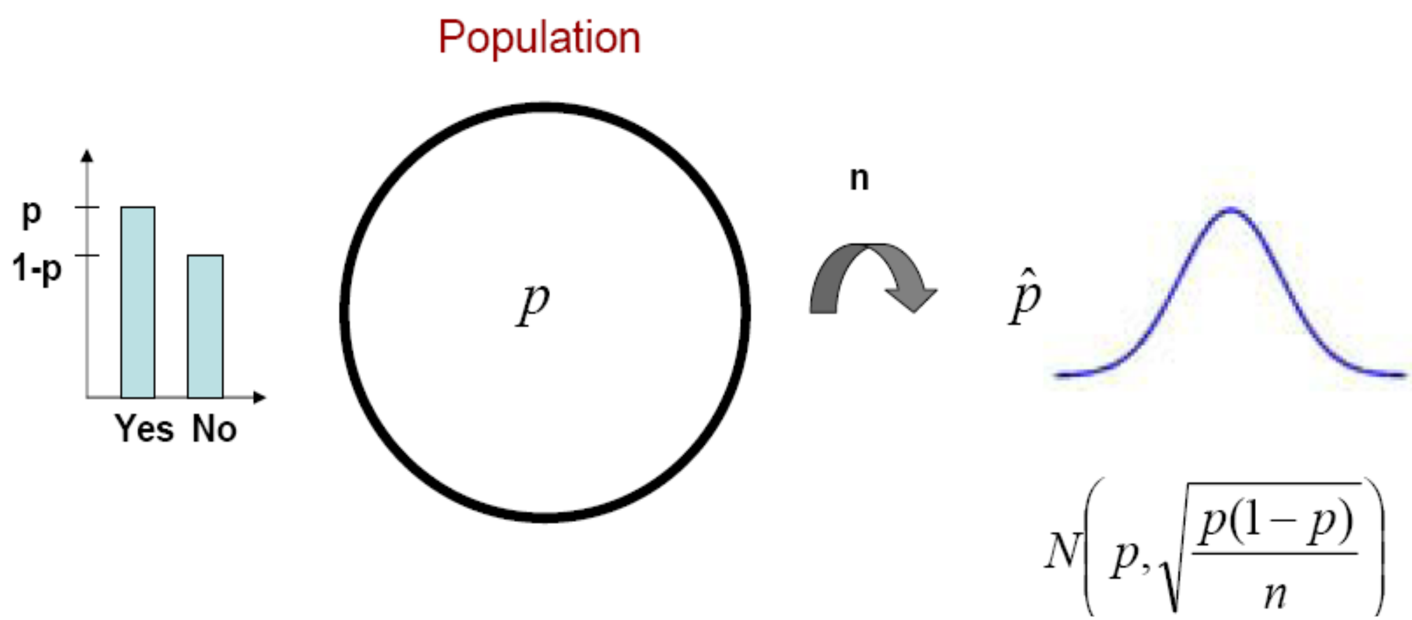
- Rule of Thumb: generally speaking, n>40 will be sufficient to justify the use of this interval. This is somewhat more conservative than the rule of thumb for the CLT, because of the additional randomness coming from $\hat{\sigma}$.

- 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*.

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

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

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

  "+" 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.

Constructing a CI

- The previous examples show the general procedure of constructing confidence intervals. Suppose X_1, X_2, …, 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

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

Three Distributions

- In statistical inference, there are three distributions playing very important roles. They are: *Chi-square distribution*, *t – distribution* and *F – distribution*.

- They arise naturally, when one deals with normal population. More specifically, each of the three distributions can be constructed by using i.i.d. normal random variables.

Chi-square

- Chi-square distribution was first brought up by K. Pearson. We often denote a Chi-square random variable with *n* degrees of freedom χ_n^2 .

- Theorem: if we have ξ_1, ξ_2, …, ξ_n i.i.d. ~ N(0,1), let

$$\eta = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2$$

then η has a χ_n^2 distribution.

- From the above theorem, if we have $\eta_1 \sim \chi_n^2$, $\eta_2 \sim \chi_m^2$, and they are independent, what is the distribution of $\eta_1+\eta_2$?

Examples

- The following are two very important examples of Chi-square distributed rv's.

Ex. If we have X_1, X_2, …, X_n i.i.d. from a normal distribution $N(\mu,\sigma^2)$, then we have

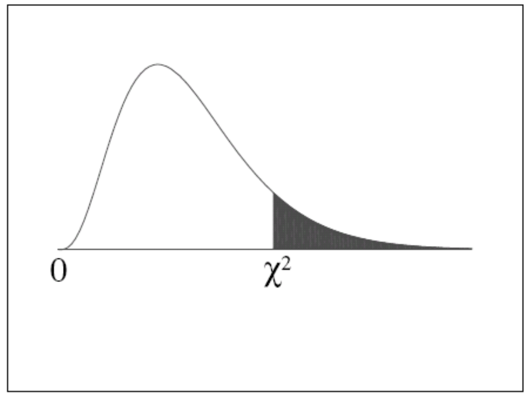
$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

Ex. If we have Y_1, Y_2, …, Y_n i.i.d. from an exponential distribution with parameter $\lambda>0$. then

$$2\lambda\sum_{i=1}^{n} Y_i \sim \chi^2_{2n}$$

- Construct CI for σ^2 and λ using the chi-square table in the appendix.

Chi-square distribution table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278

t-distribution

- t-distribution was first proposed by W. Gosset and published under the pseudonym “Student”. Thus, the t-distribution is also sometimes referred to as the “Student’s t-distribution”.

- The t-distribution density is symmetric around the origin and looks quite similar to the standard normal density. In fact, when the degree of freedom of t-distribution is large, it is indeed close to the standard norm density.

- Theorem: if $\eta_1 \sim \chi^2_n$, $\eta_2 \sim N(0,1)$, and η_1 and η_2 are independent, let

$$\zeta = \frac{\eta_2}{\sqrt{\eta_1/n}}$$

  then $\zeta \sim t_n$.

Example

- The following result is one of the most important results in statistical inference.

Ex. If we have $X_1, X_2, \ldots, X_n$ i.i.d. from a normal distribution $N(\mu,\sigma^2)$, then we have

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

- The above example also points out a very important fact that the sample mean from a normal population is independent with the sample variance (sd)!

- Construct CI for μ using the t-table in the appendix.

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^*.

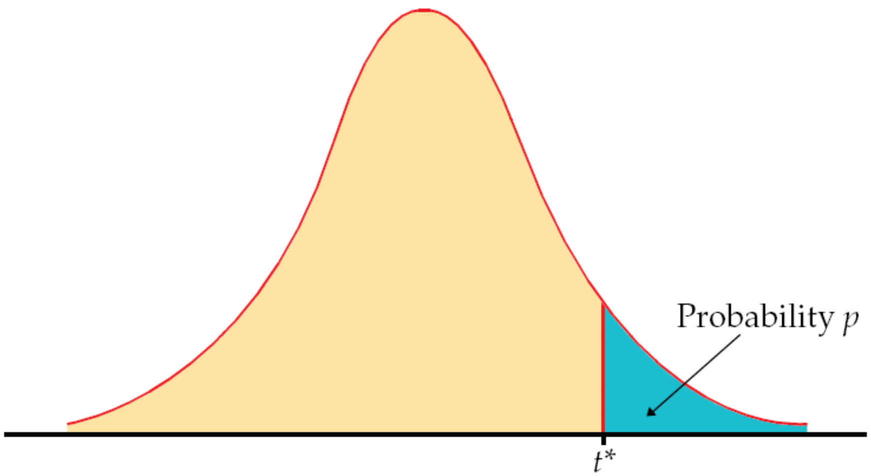


TABLE D

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

Remarks about t

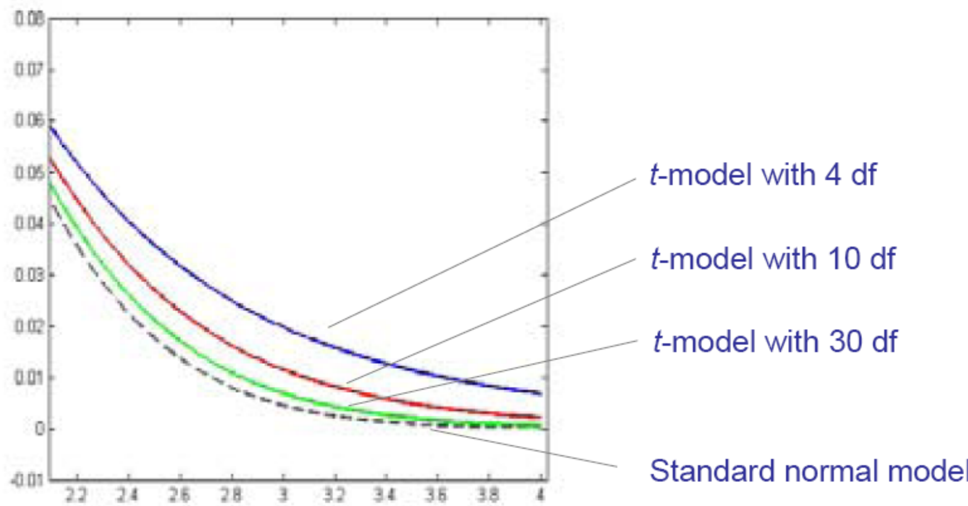
- There is a separate t-model corresponding to each degree of freedom.

- The spread of the t-model is slightly larger than that of the standard normal model. Consider the quantity

$$\frac{\sqrt{n}(\bar{\mathrm{X}} - \mu)}{\mathrm{S}}$$

  using S instead of σ introduces more variation into the statistic.

- Heavier tails:



Largest percentage changes of DJI

Largest daily percentage gains

Rank	Date	Close	Net Change	% Change
1	1933-03-15	62.10	+8.26	+15.34
2	1931-10-06	99.34	+12.86	+14.87
3	1929-10-30	258.47	+28.40	+12.34
4	1932-09-21	75.16	+7.67	+11.36
5	2008-10-13	9,387.61	+936.42	+11.08
6	2008-10-28	9,065.12	+889.35	+10.88
7	1987-10-21	2,027.85	+186.84	+10.15
8	1932-08-03	58.22	+5.06	+9.52
9	1932-02-11	78.60	+6.80	+9.47
10	1929-11-14	217.28	+18.59	+9.36
11	1931-12-18	80.69	+6.90	+9.35
12	1932-02-13	85.82	+7.22	+9.19
13	1932-05-06	59.01	+4.91	+9.08

Largest daily percentage losses

Rank	Date	Close	Net Change	% Change
1	1987-10-19	1,738.74	-508.00	-22.61
2	1929-10-28	260.64	-38.33	-12.82
3	1929-10-29	230.07	-30.57	-11.73
4	1929-11-06	232.13	-25.55	-9.92
5	1899-12-18	58.27	-5.57	-8.72
6	1932-08-12	63.11	-5.79	-8.40
7	1907-03-14	76.23	-6.89	-8.29
8	1987-10-26	1,793.93	-156.83	-8.04
9	2008-10-15	8,577.91	-733.08	-7.87
10	1933-07-21	88.71	-7.55	-7.84
11	1937-10-18	125.73	-10.57	-7.75
12	2008-12-01	8,149.09	-679.95	-7.70
13	2008-10-09	8,579.19	-678.91	-7.33

F-distribution

- The F-distribution is named after the famous statistician R.A. Fisher.

- Unlike Chi-square and t distribution, the F distribution has two degrees of freedom *n* and *m*. And these two parameters are NOT symmetric.

- Theorem: if we have $\eta_1 \sim \chi^2_n$, $\eta_2 \sim \chi^2_m$, and η_1 and η_2 are independent, let

$$\zeta = \frac{\eta_1/n}{\eta_2/m}$$

  then $\zeta \sim F_{n,m}$.

Example

Ex. Let $X_1, \ldots, X_m$ be an IID sample from a normal distribution with variance σ_1^2, let $Y_1, \ldots, Y_n$ be another IID sample from a normal distribution with variance σ_2^2. Let S_1^2 and S_2^2 denote the two sample variances, then the rv

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with v_1=m-1 and v_2=n-1.

- How to construct a confidence interval for σ_1^2 / σ_2^2?

Hypothesis Testing

- A *statistical hypothesis*, or just *hypothesis*, is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution.

- A testing problem usually contains two hypotheses: the null hypothesis, denoted by H_0, is the claim that is initially assumed to be true (the “prior belief” claim). The alternative hypothesis, denoted by H_a, is the assertion that is contradictory to H_0.

- The null hypothesis will be rejected in favor of the alternative only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the truth of the null hypothesis. The two possible conclusions from a testing analysis are then *reject* H_0 or *fail to reject* H_0.