AMS 572 Data Analysis I
Inference on one population mean μ

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Inference for μ when σ^2 is unknown, large sample for any distribution

Inference for μ when σ^2 is unknown, large sample for any distribution

Setup:

- Let $X_1, X_2, \ldots, X_n$ be a random sample for a distribution (need not be normal) with mean μ and variance σ^2.
- We assume that σ^2 is unknown.
- Sample size n is large enough $(n \geq 30)$

Theorem: Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed (i.i.d.) random variables with common mean μ and variance σ^2. Then the random variable

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

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has a limiting distribution that is normal with mean 0 and variance 1. That is,

$$Y \xrightarrow{d} Z \sim N(0,1)$$

as $n \longrightarrow \infty$

Slutsky's Theorem

- If X_n is a sequence of r.v. that converges in distribution to X, and
- Y_n is a sequence of r.v. that converges in probability to a constant c,
- then $W_n = X_n Y_n$ converges in distribution to cX
- i.e.

$$\lim_{n\to\infty} \Pr[W_n \le w] = \Pr[cX \le w]$$

Theorem: Central Limit Theorem (CLT)

To use the CLT when σ^2 unknown requires *Slutsky's Theorem*.
By both CLT and Slutsky's Theorem,

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

is approximately $N(0, 1)$.
Note: X_i's do not need to be normally distributed

CI for μ when σ^2 unknown, large sample for any distribution

- Since

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \dot{\sim} N(0,1)$$

  we have

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

- Thus $100\,(1-\alpha)\,\%$ CI for μ is given by

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

Inference for μ when σ^2 is unknown, large sample for any distribution

The derivations of the hypothesis tests (rejection region and the p-value) are almost the same as the derivation of the exact Z-test in previous set of slides.

Summary

$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$
Observed value of test statistic $Z_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$		
Rejection region : we reject H_0 in favor of H_a at the significance level α if		
$Z_0 \geq z_\alpha$	$Z_0 \leq -z_\alpha$	$\lvert Z_0\rvert \geq z_{\alpha/2}$
p-value= $P(Z_0 \geq z_0 \mid H_0)$	p-value= $P(Z_0 \leq z_0 \mid H_0)$	p-value $= P(\lvert Z_0\rvert \geq \lvert z_0\rvert \mid H_0)$ $= 2 * P(Z_0 \geq \lvert z_0\rvert \mid H_0)$
the area under $N(0,1)$ pdf to the right of z_0	the area under $N(0,1)$ pdf to the left of z_0	twice the area to the right of $\lvert z_0\rvert$

Inference for μ when σ^2 is unknown, small sample for normal distribution

Inference for μ when σ^2 is unknown, small sample for normal distribution

Setup:

- Assume that the distribution is normal
- Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal distribution with mean μ and variance σ^2. That is, $X \stackrel{iid.}{\sim} N(\mu, \sigma^2), i = 1, ..., n$.
- Assume that σ^2 is unknown.
- Assume sample size n is small.

Inference for μ when σ^2 is unknown, small sample for normal distribution

Under this scenario, the distribution of

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

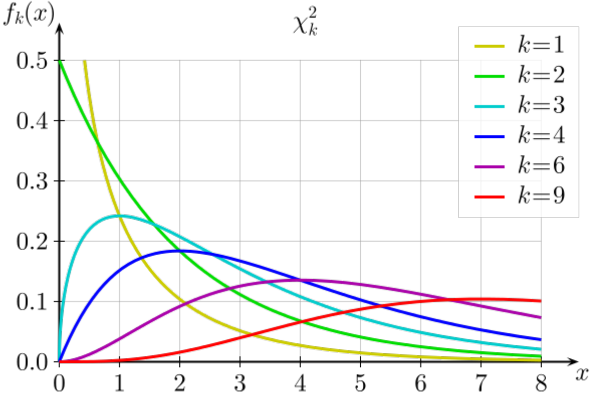
is not normal.

- If a random variable X is normally distributed with mean μ and variance σ^2, then for a random sample of size n, the quantity

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

  has a chi-square distribution with $n-1$ degrees of freedom, which we denote by χ^2_{n-1}
- Let $Z_1, Z_2, \cdots, Z_k \overset{i.i.d.}{\sim} N(0,1)$, then $W = \sum_{i=1}^k Z_i^2 \sim \chi^2_k$
- chi-square distribution is a special gamma distribution

χ^2 Distribution



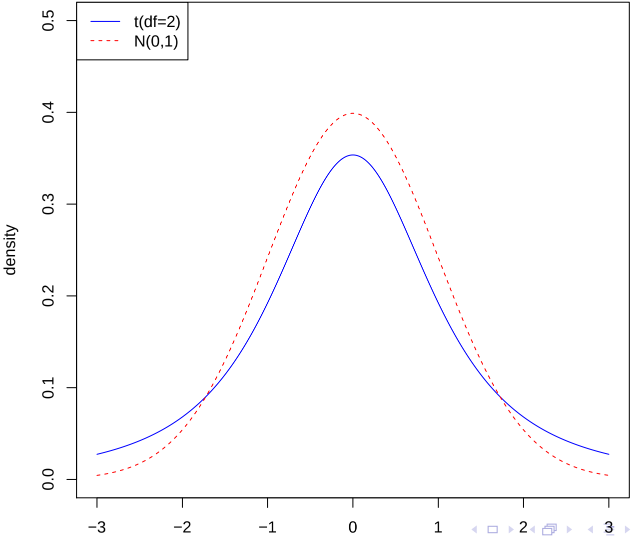
https://en.wikipedia.org/wiki/Chi-squared_distribution

- Let $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$
- If Z and W are independent, then

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

  will follow the t-distribution with ν degrees of freedom.
- t-distribution has heavier tails than normal distribution

t Distribution



Theorem: Sampling from the normal population

Let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, then

1. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
3. $\bar{X}$ and s^2 (and thus W) are independent.
   Thus we have
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

CI for μ when σ^2 unknown, small sample for normal distribution

- Since
$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$
we have $P\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X}-\mu}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$
- Thus $100\,(1-\alpha)\,\%$ CI for μ is given by

Summary

$H_0: \mu = \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Observed value of test statistic $T_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$		
Rejection region : we reject H_0 in favor of H_a at the significance level α if		
$T_0 \geq t_{n-1,\alpha}$	$T_0 \leq -t_{n-1,\alpha}$	$\|T_0\| \geq t_{n-1,\alpha/2}$
p-value= $P(T_0 \geq t_0 \| H_0)$	p-value= $P(T_0 \leq t_0 \| H_0)$	p-value= $P(\|T_0\| \geq \|t_0\| \| H_0) = 2 \cdot P(T_0\| \geq \|t_0\| \| H_0)$
the area under t_{n-1} pdf to the right of t_0	the area under t_{n-1} pdf to the left of t_0	twice the area under t_{n-1} to the right of $\|t_0\|$

Assessing Normality

- How do we assess whether the normal distribution model is a reasonable fit for a particular set of data?
- One graphical approach: quantile-quantile (QQ) plot
- Plot quantiles of the observed data distribution versus the quantiles of the normal distribution, i.e we plot the pairs

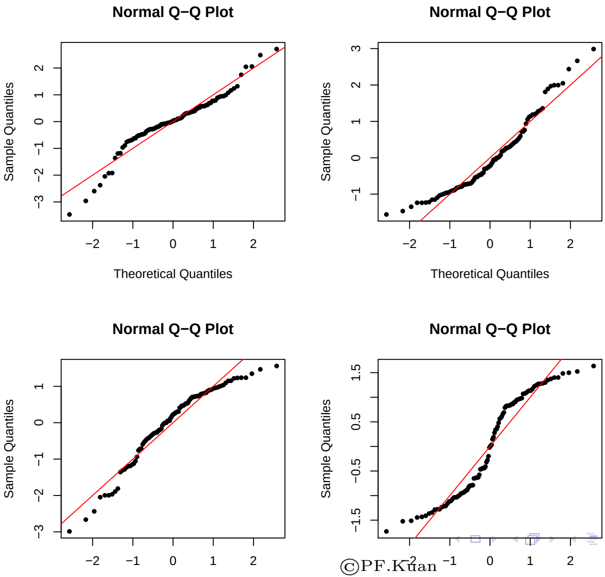
$$\left(\Phi^{-1}\left(\frac{i-0.5}{n}\right), x_{(i)}\right), i = 1, \ldots, n$$

  where $x_{(i)}$'s are the order statistics
- Straight line indicates normality assumption reasonable

QQ plot in R

```
> x <- rnorm(100)
> qqnorm(x)
> qqline(x)
```

QQ plot: Examples where normal distribution assumption is violated



Assessing Normality

- Alternatively, statistical tests for univariate normality include Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests.
- Razali et al. (2011) (Journal of Statistical Modeling and Analytics) concluded that Shapiro-Wilk test has the best power for a given significance among these tests.

Shapiro Wilk test

- Test statistic
$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
where $x_{(i)}$'s are the order statistics,
$$(a_1, \ldots, a_n) = \mathbf{m}^T V^{-1} / (\mathbf{m}^T V^{-1} V^{-1} \mathbf{m}^T)^{1/2}$$
$\mathbf{m} = (m_1, \ldots, m_n)^T$, $m_1, \ldots, m_n$ and V are are the expected values and covariance matrix of the order statistics from i.i.d $N(0,1)$, respectively.
- The empirical distribution of W is obtained via Monte Carlo simulations.

Shapiro Wilk test in R

```
> x <- rnorm(100)
> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.99273, p-value = 0.8713

> x <- rexp(100)
> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.82652, p-value = 1.763e-09
```

Example 1: Jerry is planning to purchase a sports good store. He calculated that in order to cover basic expenses, the average daily sales must be greater than \$525.

Scenario A. He checked the daily sales of 36 randomly selected business days, and found the average daily sales to be \$565 with a standard deviation of \$150.

Scenario B. Now suppose he is only allowed to sample 9 days. And the 9 days sales are \$510, 537, 548, 592, 503, 490, 601, 499, 640.

For A and B, determine if Jerry can conclude the daily sales to be greater than \$525 at the significance level of $\alpha = 0.05$. What is the p-value for each scenario?

Example 2 (Recap): Let $X_1, X_2, \ldots, X_n$ be i.i.d random sample of size n ($n \geq 30$) from a population with unknown and non-normal distribution. Derive the $100\,(1-\alpha)\,\%$ CI for μ.

Ans: According to the CLT and Slutsky's Theorem,
$Z = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \dot{\sim} N\,(0,1)$,
where $s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}}$.
Since $P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$,
then the $100\,(1-\alpha)\,\%$ CI for μ is $(\overline{X} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}})$.
*If the population variance σ^2 is known, then the CI will be

$$(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}).$$

Example 3 (Recap): Let $X_1, X_2, \ldots, X_n$ be i.i.d random sample of size n from $N(\mu,\ \sigma^2)$, where σ^2 is unknown. Derive the $100\,(1-\alpha)\,\%$ CI for μ.

Ans: Since $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$, then
$P\left(-t_{n-1,\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \leq t_{n-1,\frac{\alpha}{2}}\right) = 1-\alpha,$
so the $100\,(1-\alpha)\,\%$ CI for μ is

$$(\overline{X} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}).$$