In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
from IPython.core.pylabtools import figsize # import figsize
#figsize(12.5, 4) # 设置 figsize
from scipy.stats import chi2
from scipy.stats import t
from scipy.stats import f
from scipy.stats import norm
```

Distributions and Fisher Tests

Normal Distribution

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Chi2 Distribution

$(n-1)S^2/\sigma^2$ is chi-squared distributed with $n-1$ degrees of freedom

20.4. Chi-Squared Test. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution and let S^2 denote the sample variance. Let σ^2 be the unknown population variance and σ_0^2 a null value of that variance. Then a test for the variance based on the statistic

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma_0^2}$$

is called a ***chi-squared test***. We reject at significance level α

- $H_0 : \sigma = \sigma_0$ if $\chi^2_{n-1} > \chi^2_{\alpha/2,n-1}$ or $\chi^2_{n-1} < \chi^2_{1-\alpha/2,n-1}$,
- $H_0 : \sigma \le \sigma_0$ if $\chi^2_{n-1} > \chi^2_{\alpha,n-1}$,
- $H_0 : \sigma \ge \sigma_0$ if $\chi^2_{n-1} < \chi^2_{1-\alpha,n-1}$.

T Distribution

random variable

$$T_\gamma = \frac{Z}{\sqrt{\chi^2_\gamma/\gamma}}$$

is said to follow a T-distribution with γ degrees of freedom.

$$T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

20.1. T-Test. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution and let $\overline{X}$ denote the sample mean, S^2 the sample variance. Let μ be the unknown population mean and μ_0 a null value of that mean. Then any test based on the statistic

$$T_{n-1} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

is called a ***T-test***.

We reject at significance level α

- $H_0\colon \mu = \mu_0$ if $|T_{n-1}| > t_{\alpha/2,n-1}$,
- $H_0\colon \mu \leq \mu_0$ if $T_{n-1} > t_{\alpha,n-1}$,
- $H_0\colon \mu \geq \mu_0$ if $T_{n-1} < -t_{\alpha,n-1}$.

F Distribution

$$F_{\gamma_1,\gamma_2} = \frac{X^2_{\gamma_1}/\gamma_1}{X^2_{\gamma_2}/\gamma_2}$$

$$F_{n_1-1,n_2-1} = \frac{[(n_1-1)S_1^2/\sigma_1^2]/(n_1-1)}{[(n_2-1)S_2^2/\sigma_2^2]/(n_2-1)} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

23.5. F-Test. Let S_1^2 and S_2^2 be sample variances based on independent random samples of sizes n_1 and n_2 drawn from normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. Then a test based on the statistic

$$F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2}$$

is called an ***F-test***.

We reject at significance level α

- $H_0\colon \sigma_1 \leq \sigma_2$ if $\dfrac{S_1^2}{S_2^2} > f_{\alpha,n_1-1,n_2-1}$,
- $H_0\colon \sigma_1 \geq \sigma_2$ if $\dfrac{S_2^2}{S_1^2} > f_{\alpha,n_2-1,n_1-1}$,
- $H_0\colon \sigma_1 = \sigma_2$ if $\dfrac{S_1^2}{S_2^2} > f_{\alpha/2,n_1-1,n_2-1}$ or $\dfrac{S_2^2}{S_1^2} > f_{\alpha/2,n_2-1,n_1-1}$

In [2]:
```python
# chi square distribution
#percents = [0.995, 0.990, 0.975, 0.950, 0.900, 0.100, 0.050, 0.025, 0.010, 0.005]
#print(np.array([chi2.isf(percents, df=i) for i in range(1, 47)]))
# t distribution
#percents = [0.100, 0.050, 0.025, 0.010, 0.005, 0.001, 0.0005]
#print(np.array([t.isf(percents, df=i) for i in range(1, 46)]))
# F distribution
#alpha = 0.2
#print(np.array([f.isf(alpha, df1, df2) for df1 in range(1, 11) for df2 in range(1, 16)]).reshape(10, -1).T)
# normal distribution
#print(norm.ppf(np.arange(0, 0.99, 0.001).reshape(-1, 10)))

def statistic(x):
    print(x)
    x = np.array(x).reshape(-1,1)
    mean = x.mean()
    n = x.shape[0]
    s = np.sqrt(np.sum((x-x.mean())**2)/(n-1))
    s_square = s**2
    cache =  { 'n':n
             , 'mean':mean
             , 's_square':s_square
             , 's':s

             }
    print(cache)
    return cache
```

In [3]:
```python
x = [708, 732, 731, 677, 748, 702, 696, 692, 716, 729,697, 681, 704, 740, 710, 687, 731, 704, 702, 698]
x_dict = statistic(x)
```

```
[708, 732, 731, 677, 748, 702, 696, 692, 716, 729, 697, 681, 704, 740, 710, 687, 731, 704, 702, 698]
{'n': 20, 'mean': 709.25, 's_square': 399.5657894736842, 's': 19.98914178932363}
```

In [4]:
```python
# normal distribution
# norm.ppf(percent)

# chi square distribution
# chi2.isf(percent,df)

# T distribution
# t.isf(percent, df)

# F distribution
# f.isf(alpha,df1,df2)
```

Confidences for Different Estimators

Confidence for Mean

21.1. Example. The confidence interval for the mean derived previously has the form

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},$$

Confidence for Variance

16.13. Theorem. Let $X_1, \dots, X_n$, $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2. A $100(1-\alpha)\%$ confidence interval on σ^2 is given by

$$\left[(n-1)S^2/\chi^2_{\alpha/2,n-1},\ (n-1)S^2/\chi^2_{1-\alpha/2,n-1}\right].$$

Non-Paramatic Test

Sign Test

21.2. Sign Test. Let $X_1, \dots, X_n$ be a random sample of size n from an arbitrary continuous distribution and let

$$Q_+ = \#\{X_k : X_k - M_0 > 0\}, \qquad Q_- = \#\{X_k : X_k - M_0 < 0\}.$$

We reject at significance level α

- $H_0 : M \le M_0$ if $P[Q_- \le k \mid M = M_0] < \alpha$,
- $H_0 : M \ge M_0$ if $P[Q_+ \le k \mid M = M_0] < \alpha$,
- $H_0 : M = M_0$ if $P[\min(Q_-, Q_+) \le k \mid M = M_0] < \alpha/2$.

Wilcoxon Signed Rank Test

21.4. Wilcoxon Signed Rank Test. Let $X_1, \dots, X_n$ be a random sample of size n from a symmetric distribution. Order the n absolute differences $|X_i - M|$ according to magnitude, so that $X_{R_i} - M_0$ is the R_ith smallest difference by modulus. If ties in the rank occur, the mean of the ranks is assigned to all equal values.

Let

$$W_+ = \sum_{R_i > 0} R_i, \qquad |W_-| = \sum_{R_i < 0} |R_i|.$$

We reject at significance level α

- $H_0 : M \le M_0$ if W_- is smaller than the critical value for α,
- $H_0 : M \ge M_0$ if W_+ is smaller than the critical value for α,
- $H_0 : M = M_0$ if $W = \min(W_+, |W_-|)$ is smaller than the critical value for $\alpha/2$.

The distribution of the test statistics is complicated; there are tables that give critical values for small sample sizes, typically up to $n \le 20$.

For non-small sample sizes ($n \ge 10$) a normal distribution with parameters

$$\mathrm{E}[W] = \frac{n(n+1)}{4}, \qquad \mathrm{Var}[W] = \frac{n(n+1)(2n+1)}{24}.$$

may be used as an approximation. However, in that case the variance needs to be reduced if there are ties: for each group of t ties, the variance is reduced by $(t^3 - t)/48$.

Pooled Tests: Comparing Means

Comparing Means with Known Variance

A Point Estimator for the Difference of Means

We take random samples $\overline{X}^{(1)}$ and $\overline{X}^{(2)}$ of sizes n_1 and n_2 from the populations, we can find a point estimator for the difference of the two means

$$\widehat{\mu_1 - \mu_2} := \widehat{\mu}_1 - \widehat{\mu}_2 = \overline{X}^{(1)} - \overline{X}^{(2)}.$$

Since

$$\overline{X}^{(1)} \sim N(\mu_1, \sigma_1^2/n_1), \qquad \overline{X}^{(2)} \sim N(\mu_2, \sigma_2^2/n_2),$$

we see that $\overline{X}_1 - \overline{X}_2$ is normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$, i.e.,

$$\frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is a standard normal random variable.

$$\mu_1 - \mu_2 = \overline{x}^{(1)} - \overline{x}^{(2)} \pm z_\alpha \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Comparing Means with Equal Variance: Student-T Test

Now suppose that the variances are equal but unknown,

$$\sigma_1^2 = \sigma_2^2 =: \sigma^2.$$

Then

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}.$$

is standard normal

Similarly to (22.1), we define the ***pooled estimator for the variance***

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{24.1}$$

$$
\begin{aligned}
T_{n_1+n_2-2} &= \frac{Z}{\sqrt{X^2_{n_1+n_2-2}/(n_1+n_2-2)}} \\
&= \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2(1/n_1+1/n_2)}}
\end{aligned}
$$

$$
(\overline{X}_1-\overline{X}_2) \pm t_{\alpha/2,n_1+n_2-2}\sqrt{S_p^2(1/n_1+1/n_2)},
$$

Comparing Means with Unequal Variance: Welch-Satterthwaite Approximation

We are interested in the case $k=2$, $\lambda_1 = 1/n_1$ and $\lambda_2 = 1/n_2$. Then

$$
\gamma = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}. \tag{24.2}
$$

and

$$
\gamma \cdot \frac{S_1^2/n_1 + S_2^2/n_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}
$$

follows approximately a chi-squared distribution with γ degrees of freedom. It is then easy to see that

$$
T_\gamma = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}
$$

follows a T-distribution with γ degrees of freedom.

$$
T_\gamma = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}
$$

is called a ***Welch's (pooled) test for equality of means***. We reject at significance level α

- $H_0: \mu_1-\mu_2 = (\mu_1-\mu_2)_0$ if $|T_\gamma| > t_{\alpha/2,\gamma}$,
- $H_0: \mu_1-\mu_2 \leq (\mu_1-\mu_2)_0$ if $T_\gamma > t_{\alpha,\gamma}$,
- $H_0: \mu_1-\mu_2 \geq (\mu_1-\mu_2)_0$ if $T_\gamma < -t_{\alpha,\gamma}$.

Paired Test: Comparing Means

Wilcoxon Rank-Sum Test

25.1. Wilcoxon Rank-Sum Test. Let X and Y be two random samples following some continuous distributions.

Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$, $m \le n$, be random samples from X and Y and associate the rank R_i, $i = 1, \dots, m+n$, to the R_ith smallest among the $m+n$ total observations. If ties in the rank occur, the mean of the ranks is assigned to all equal values.

Then the test based on the statistic

$$W_m := \text{sum of the ranks of } X_1, \dots, X_m.$$

is called the ***Wilcoxon rank-sum test***.

We reject $H_0\colon P[X > Y] = 1/2$ (and similarly the analogous one-sided hypotheses) at significance level α if W_m falls into the corresponding critical region.

For large values of m ($m \ge 20$), W_m is approximately normally distributed with

$$\mathrm{E}[W_m] = \frac{m(m+n+1)}{2}, \qquad \mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12}.$$

If there are many ties, the variance may be corrected by taking

$$\mathrm{Var}[W_m] = \frac{mn(m+n+1)}{12 - \sum\limits_{\text{groups}} \frac{t^3+t}{12}}$$

where the sum is taken over all groups of t ties. However, the best way to deal with ties is still a topic of current research.

Example:

$$\begin{aligned} w_{14} &= 1.5 + 4 + 14.5 + 20 + 22 + 22 + 26 \\ &\quad + 31 + 33 + 37 + 41 + 41 + 43.5 + 43.5 \\ &= 380 \end{aligned}$$

Given the large sample sizes, we use a normal approximation for the test statistic (most tables only include values for $m, n \le 20$). We have

$$\begin{aligned} \mathrm{E}[W_{14}] &= \frac{14(14+30+1)}{2} = 315, \\ \mathrm{Var}\, W_{14} &= \frac{14 \cdot 30(14+30+1)}{12} = 1575 \end{aligned}$$

Therefore,

$$Z = \frac{W_m - 315}{\sqrt{1575}}$$

follows a standard normal distribution if $P[X_{\text{undergrad}} > X_{\text{grad}}] = 1/2$. The value of our test statistic is

$$z = \frac{380 - 315}{\sqrt{1575}} = 1.64.$$

Using the normal distribution table, we find a P-value of

$$P[Z \geq 1.64] = 0.0505.$$

Paired-T Test:

D = X - Y

Then

$$T_{n-1} = \frac{\overline{D} - \mu_D}{\sqrt{S_D^2/n}}$$

follows a T-distribution with $n - 1$ degrees of freedom.

Chi2 Goodness Fit

1D

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

will follow a chi-squared distribution with $k - 1 - m$ degrees of freedom, where m is the number of parameters that we estimate.

Cochran's Rule states that wee should require

$$\begin{aligned} \mathrm{E}[X_i] &= np_i \geq 1 && \text{for all } i = 1, \ldots, k, \\ \mathrm{E}[X_i] &= np_i \geq 5 && \text{for 80\% of all } i = 1, \ldots, k, \end{aligned}$$

Example: Poisson Distribution

From Example 15.5 we know that a maximum-likelihood estimator for k is the sample mean,

$$\widehat{k} = \overline{X} = \frac{1}{60}(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3) = 0.75.$$

In order to apply the multinomial distribution, we first calculate

$$\begin{aligned}
P[X = 0] &= \frac{e^{-\widehat{k}}\widehat{k}^0}{0!} = 0.472 \\
P[X = 1] &= \frac{e^{-\widehat{k}}\widehat{k}^1}{1!} = 0.354 \\
P[X = 2] &= \frac{e^{-\widehat{k}}\widehat{k}^2}{2!} = 0.133 \\
P[X \geq 3] &= 1 - P[X = 0] - P[X = 1] - P[X = 2] = 0.041
\end{aligned}$$

2D

From Example 15.5 we know that a maximum-likelihood estimator for k is the sample mean,

$$\widehat{k} = \overline{X} = \frac{1}{60}(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3) = 0.75.$$

In order to apply the multinomial distribution, we first calculate

$$\begin{aligned}
P[X = 0] &= \frac{e^{-\widehat{k}}\widehat{k}^0}{0!} = 0.472 \\
P[X = 1] &= \frac{e^{-\widehat{k}}\widehat{k}^1}{1!} = 0.354 \\
P[X = 2] &= \frac{e^{-\widehat{k}}\widehat{k}^2}{2!} = 0.133 \\
P[X \geq 3] &= 1 - P[X = 0] - P[X = 1] - P[X = 2] = 0.041
\end{aligned}$$