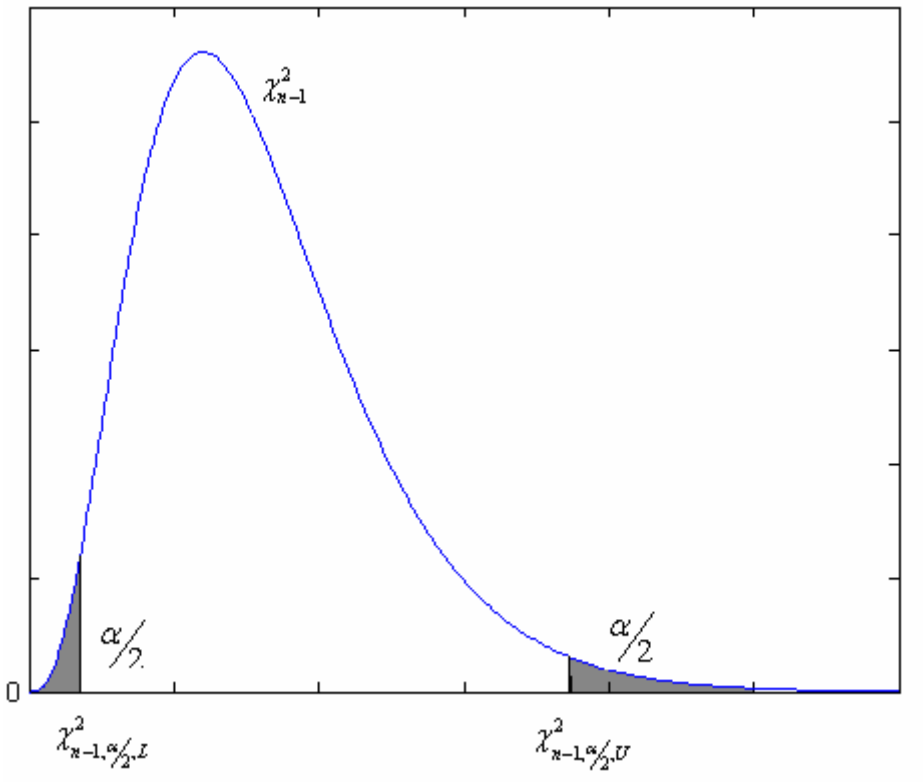
Introduction to mathematical Statistics

1. Inference on one population variance σ^2, population is normal (§6.4)

1） Point estimator: $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ and s^2 is unbiased estimator of σ^2

Pivotal Quantity for the inference on σ^2 (P.Q.): $W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

2） Confidence Interval for σ^2:



$$P(\chi^2_{n-1,L,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,U,\alpha/2}) = 1-\alpha$$

$$P\left(\frac{n-1}{\chi^2_{U,\alpha/2}} \le \frac{\sigma^2}{s^2} \le \frac{n-1}{\chi^2_{L,\alpha/2}}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{U,\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{L,\alpha/2}}\right) = 1-\alpha$$

Hence, the $100(1-\alpha)\%$ confidence interval for σ^2 is $\left[\frac{(n-1)s^2}{\chi^2_{n-1,U,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,L,\alpha/2}}\right]$

Example 1. Home buyers can choose a variety of ways to finance mortgages, ranging from fixed-rate thirty-year notes to one-year adjustable, where interest rates can move up or down from year to year. We sampled n=9 lenders for one-year adjustable loan and the sample standard deviation for those nine interest rates is s=0.22%. Please construct a 95% confidence interval for such interest assuming the population is normal.

Solution: This is inference on one population variance, normal population.

n=9 s=0.22% α=0.05 $\chi^2_{8,U,0.025} = 17.535$ $\chi^2_{8,L,0.025} = 2.18$

The 95% CI for σ^2 in the unit of $(\%)^2$ is: $\left[\frac{8*0.22^2}{17.535}, \frac{8*0.22^2}{2.18}\right] \approx [0.022, 0.178]$

2. Two-sample problems (*This is §6.4 except for the inference on two populations means when the two samples are paired.)

2.1: Inference on two population means, paired samples: *this reduces to the inference on one population mean based on the paired differences*

Paired samples approach

	Population 1 [Sample 1]	Population 2 [Sample 2]	Paired difference
Pair 1	150	140	10
Pair 2	137	167	-30
Pair 3	172	155	17
…	…	…	…

Example 2. Do fraternities help or hurt your academic progress at college? To investigate this question, 5 students who joined fraternities in 1998 were randomly selected. It was shown that their GPA before and after they joined the fraternities are as follows. Please construct a 95%

confidence interval for the mean changes in the GPA's for students who had joined the fraternities.

Student	1	2	3	4	5
Before	3	4	3	3	2
After	2	3	3	2	1
Difference	1	1	0	1	1

Solution:

Since the sample size is small, we assume the difference follows a normal distribution.

$\overline{X_d} = 0.8, S_d = 0.447, n = 5, \alpha = 0.05$; $\mu_d = \mu_{before} - \mu_{after}$

Pivotal quantity : $T = \dfrac{\overline{X_d} - \mu_d}{S_d / \sqrt{n}} \sim t_{n-1}$

The 100(1-α)% CI for μ_d is $\left[\overline{X_d} - t_{n-1,\alpha/2} * S_d / \sqrt{n},\ \overline{X_d} + t_{n-1,\alpha/2} * S_d / \sqrt{n}\right]$

For the given problem, we wish to construct a 95% CI, and we have $t_{4,0.025} = 2.776$, and the corresponding CI is: $[0.245,\ 1.355]$

Since zero is not included in this interval, we claim that we are 95% sure that the mean GPA has decreased after the students joined the fraternities.

2.2: Inference on two population variances, independent samples: *Both samples are from normal populations.*

"Unpaired samples" [Independent Samples]

Population 1 (Men) [Sample 1]	Population 2 (Women) [Sample 2]
X_1	Y_1
X_2	Y_2
…	…
X_{n_1}	Y_{n_2}

Data: $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

$Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

① **Parameter of interest:** $\frac{\sigma_1^2}{\sigma_2^2}$

② **Point estimator:** $\hat{(\frac{\sigma_1^2}{\sigma_2^2})} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{S_1^2}{S_2^2}$

③ **P.Q:**

***Definition*:** F-Distribution

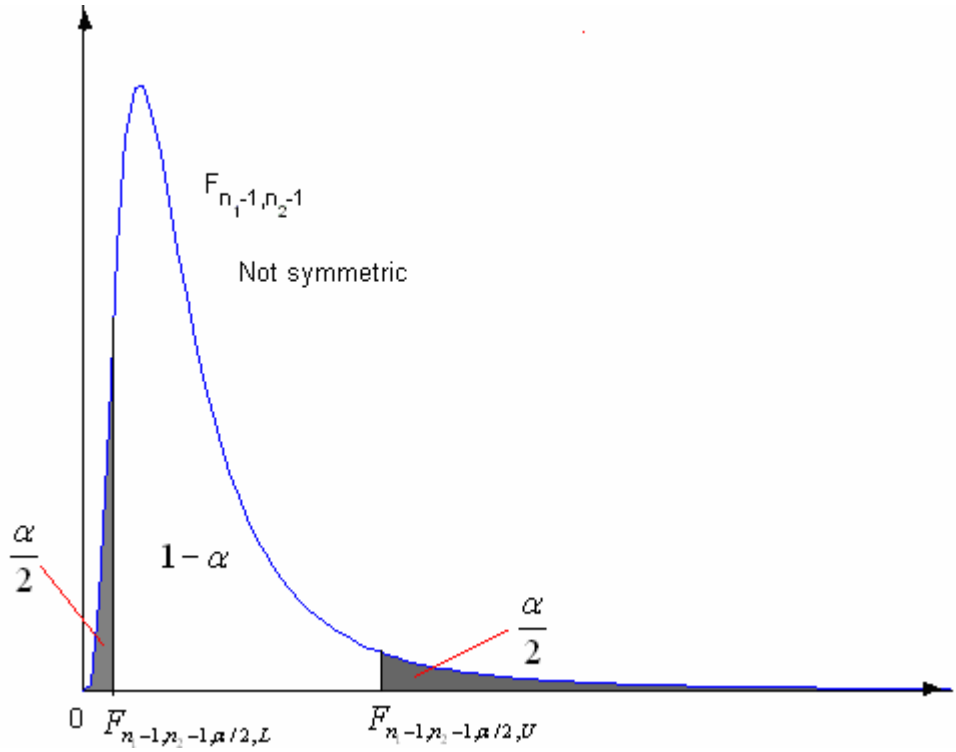
Let $W_1 \sim \chi^2_{k_1}, W_2 \sim \chi^2_{k_2}$, and W_1, W_2 are independent. Then $F = \frac{W_1/k_1}{W_2/k_2} \sim F_{k_1,k_2}$

$\because \quad \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$

$\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$

$\therefore \quad F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2} \Big/ (n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2} \Big/ (n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1} \rightarrow$ Pivotal Quantity

④ **Conference Interval for** $\frac{\sigma_1^2}{\sigma_2^2}$



$$1-\alpha = P(F_{n_1-1,n_2-1,\alpha/2,L} \le F \le F_{n_1-1,n_2-1,\alpha/2,U}) = P(F_L \le \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \le F_U)$$

$$= P(\frac{\frac{S_1^2}{S_2^2}}{F_U} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\frac{S_1^2}{S_2^2}}{F_L})$$

Therefore the 100(1-α)% CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is $\left[\frac{\frac{S_1^2}{S_2^2}}{F_U}, \frac{\frac{S_1^2}{S_2^2}}{F_L}\right]$

*** A trick for the F distribution**

If $F \sim F_{k_1,k_2}$**, then** $\frac{1}{F} \sim F_{k_2,k_1}$

Some F-table only gives the upper bound. If we know $F_{k_1,k_2,\alpha,U}$ **, how to find** $F_{k_1,k_2,\alpha,L}$**?**

$$\alpha = P(F \le F_{k_1,k_2,\alpha,L}) = P(\frac{1}{F} \ge F_{k_2,k_1,\alpha,U}) = P(\frac{1}{F} \ge \frac{1}{F_{k_1,k_2,\alpha,L}})$$

$$\because \boxed{\frac{1}{F_{k_1,k_2,\alpha,L}} = F_{k_2,k_1,\alpha,U}}$$

2.3: Inference on two population means, independent samples: *Pooled variance t: Both samples are from normal populations. Furthermore, we assume the population variances are unknown but equal, that is:* $\boxed{\sigma_1^2 = \sigma_2^2 = \sigma^2}$

1) Parameter of interest

$\mu_1 - \mu_2$ (or $\frac{\mu_1}{\mu_2}$)

2) Point estimator for the parameter of interest

$\overline{X} - \overline{Y}$ (or $\frac{\overline{X}}{\overline{Y}}$)

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

The ratio is not used because it is very hard to figure out the point estimator.

3) $$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Z is not a pivotal quantity for $(\mu_1 - \mu_2)$ since σ is unknown.

4) $$\left.\begin{aligned} W_1 &= \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1} \\ W_2 &= \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1} \end{aligned}\right\} \textit{independent}$$

$$W = W_1 + W_2 \sim \chi^2_{n_1 + n_2 - 2}$$

5) W_1, W_2, $\overline{X}$, and $\overline{Y}$ are independent.

Thus, W and Z are independent.

$$T = \frac{Z}{\sqrt{\frac{W}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

$$= \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

S_p^2 is called the pooled-variance.

Therefore the pivotal quantity is $\boxed{T = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}}$

6) $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$:

$$\boxed{\overline{X}-\overline{Y} \pm t_{n_1+n_2-2,\alpha/2} \cdot S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Example 3. A new method of making concrete blocks has been proposed. To see whether or not the new method increases the compressive strength, 5 sample blocks are made by each method.

New Method	14	15	13	15	16
Old Method	13	15	13	12	14

a. Construct a 95% CI for the ratio of the variances of the 2 methods.

b. Construct a 95% CI for the mean difference of the 2 methods.

Solution:

$n_1 = 5$, $\overline{X} = 14.6$, $S_1^2 = 1.3$, $n_2 = 5$, $\overline{Y} = 13.4$, $S_2^2 = 1.3$

a. Assume both populations are normal, first we construct the 95% CI for σ_1^2/σ_2^2

$F_{4,4,0.025,U} = 9.60$, $F_{4,4,0.025,L} = 1/F_{4,4,0.025,U} = 1/9.60 \approx 0.104$

Therefore the 95% CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is $\left[\frac{1}{F_U}, \frac{1}{F_L}\right] \approx [0.104, 9.60]$

Since 1 $\left(\frac{\sigma_1^2}{\sigma_2^2} = 1\right)$ is inside the interval, it is thus reasonable to assume that $\sigma_1^2 = \sigma_2^2$ and we can use the pooled variance t for part b.

b. The 95% CI for $(\mu_1 - \mu_2)$ is $(\overline{X_1} - \overline{X_2}) \pm t_{n_1+n_2-2,0.025} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

That is, $(14.6 - 13.4) \pm 2.306 \cdot 1.14 \cdot \sqrt{\frac{1}{5} + \frac{1}{5}}$

That is: $[-0.46, 2.86]$

Since 0 $(\mu_1 - \mu_2 = 0)$ is included in the interval, we can not conclude that the new method has more compression strength than the old method.

2.4: Inference on two population means, independent samples: *normal populations, both variances are known that is:* σ_1^2, σ_2^2 *are both known*

1) Parameter of interest

   $\mu_1 - \mu_2$

2) Point estimator for the parameter of interest

   $\overline{X} - \overline{Y}$

   $\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

3) $Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

   Z is a pivotal quantity for $(\mu_1 - \mu_2)$ since σ_1^2, σ_2^2 are known.

4) $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$:

$$\overline{X}-\overline{Y}\pm Z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

2.5: Large sample inference on two population means, independent samples, both samples are large: *any populations, variances either known or unknown – based on the central limit theorem (CLT)*

1) Parameter of interest

$\mu_1-\mu_2$

2) Point estimator for the parameter of interest

$\overline{X}-\overline{Y}$

$$\overline{X}-\overline{Y}\dot{\sim}N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

3) $$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\dot{\sim}N(0,1)$$

Z is a pivotal quantity for $(\mu_1-\mu_2)$ when σ_1^2,σ_2^2 are known.

$$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\dot{\sim}N(0,1)$$

Z is a pivotal quantity for $(\mu_1-\mu_2)$ when σ_1^2,σ_2^2 are unknown.

4) $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$:

$$\overline{X}-\overline{Y}\pm Z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$\overline{X}-\overline{Y}\pm Z_{\alpha/2}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

2.6: Large sample inference on two population proportions, independent samples, both samples are large:– *based on the central limit theorem (CLT)*

Sample 1 (population proportion -- π_1): n_1, X_1

Sample 2 (population proportion -- π_2): n_2, X_2

1). Derivation of the Pivotal Quantity:

Parameter of Interest: $\pi_1 - \pi_2$

$$\hat{\pi}_1 = \frac{X_1}{n_1} \sim N(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1})$$
$$\hat{\pi}_2 = \frac{X_2}{n_2} \sim N(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2})$$

Point Estimator of $(\pi_1 - \pi_2)$ is: $(\hat{\pi}_1 - \hat{\pi}_2)$

Pivotal Quantity (P.Q.) for $(\pi_1 - \pi_2)$:

$$\hat{\pi}_1 - \hat{\pi}_2 \sim N(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2})$$

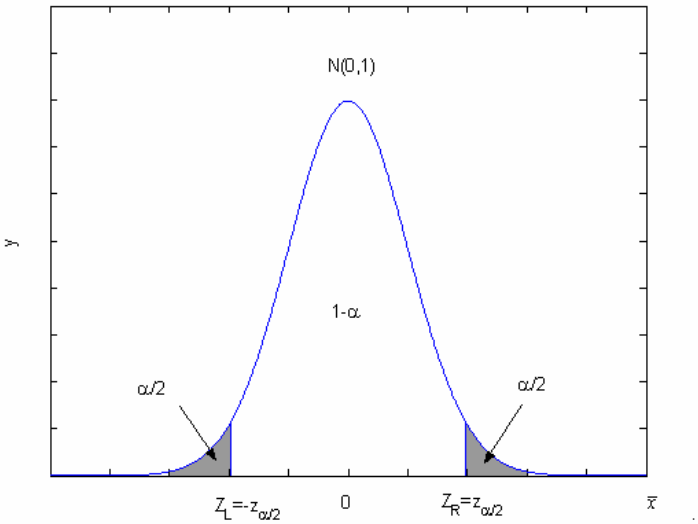
$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

The above is not a P.Q. because π_1, π_2 are used in $\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$

However the following is a P.Q.:

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \sim N(0,1)$$

2. The 100(1-α)% CI for $\pi_1 - \pi_2$



$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \approx 1-\alpha$$

… We can follow the same procedure as before to arrive at the following formula for the **100(1-α)% CI for** $\pi_1 - \pi_2$:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Example 4: We are interested in testing whether taking vitamin C would reduce the incidence of cold. A sample of 139 subjects was asked to take the VC pills continuously for 2 winter months and the incidence of cold is found to be 17 during this period. A sample of 140 subjects comparable in age, gender and health conditions was asked to take the placebo (pills that look just like the VC pills) for the same period of time and 31 subjects in this group caught cold during the same 2-month period. Please conduct a 95% confidence interval for the difference of these incidence rates.

Solution:
Both samples are large since 17 > 5 and 139-17 >5; 31 > 5 and 140-31 > 5.

For the vitamin C group, the sample proportion catching cold is $\hat{p}_1 = 17/139 = 0.122$.

For the placebo group, the sample proportion catching cold is $\hat{p}_2 = 31/140 = 0.221$.

The 95% confidence interval for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm Z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

That is: [-0.187, -0.011]; we therefore conclude that VC does help reduce cold incidence rate.

Quiz 7. Take home quiz. It is basically Homework 5 with problems from sections 6.4 and 6.5 of our textbook (see below). Two problems will be chosen to be graded. It is due before class on Monday, April 4. (No quiz will be given on Friday, April 1.)

Homework 5 (Quiz 7):
6.4.2, 6.4.4, 6.4.5, 6.4.11, 6.5.1, 6.5.2, 6.5.3, 6.5.5, 6.5.6, 6.5.8, 6.5.13, 6.5.14