Data Analysis

Spring Semester, 2023
February 7, 2023
Lecture 5

Feb 23: ***Examination One: Chapters 3, 4, 5, 6, and 7***

Chapter 4, Probability and Probability Distributions

Definition of variance of the random variable X: $var(X) = E((X - EX)^2)$

Important identity:

$$var(X) = E((X - EX)^2) = E(X^2) - (EX)^2$$

Chapter Five,
Inferences about Population Central Values

Distribution of Sample Mean:

- Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from Y which has the distribution $N(\mu, \sigma^2)$.

  THEN the distribution of $\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n}$, the sample mean, is $N(\mu, \sigma^2/n)$.
- Central Limit Theorem (CLT): When $Y_1, Y_2, \ldots, Y_n$ is a random sample of size n from Y which has expected value μ and variance $\sigma^2 < \infty$, then the distribution of $\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n}$ is asymptotically $N(\mu, \sigma^2/n)$. For a random sample of size n, it is always true that $E(\bar{Y}_n) = \mu$ and $\mathrm{var}(\bar{Y}_n) = \frac{\sigma^2}{n}$. The CLT allows probability calculations that increase in accuracy as the sample size increases.

Testing a Statistical Hypothesis (Variance Unknown).

- We cannot put $\bar{Y}_n$ in standard score form: $Z = \frac{\bar{Y}_n - \mu_0}{\sigma/\sqrt{n}}$ because we do not know σ.

- Instead, we use an estimate of $\sigma^2, \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1}$, which has $n-1$ degrees of freedom.
- We put $\bar{Y}_n$ in studentized standard score form: $T_{n-1} = \frac{\bar{Y}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}$.
- Student showed that the percentiles from the standard normal (here 1.645, 1.960, and 2.576) had to be stretched. The amount of stretching is determined by the number of degrees of freedom in $\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1}$; here, $n-1$degrees of freedom. These values are tabulated and will be given to you in your examinations.

Confidence Interval, Variance Unknown

- As always, we use the data $S^2$to estimate σ^2 and stretch our normal percentile (here 2.576) using the degrees of freedom of S^2. The stretch of 2.576 is 3.499.
- The 99% confidence interval for *E(Y)* is $\bar{y}_n \pm t_{n-1,2.576}\frac{\hat{\sigma}}{\sqrt{n}}$.
- In the example problem, the 99% confidence interval for *E(Y)* is $\bar{y}_8 \pm t_{7,2.576}\frac{\hat{\sigma}}{\sqrt{n}} = 43.2 \pm 3.499\sqrt{\frac{517.5}{8}} = 43.2 \pm 3.499(8.04) = 43.2 \pm 28.1$.

Chapter Six
Inferences Comparing Two Population Central Values

New Problem (just slightly different from Chapter 4, number 7):
The random variables $W_1$and W_2 are independent. Their expected values are $E(W_1) = \mu_1$ and $E(W_2) = \mu_2$. Their variances are $\text{var}(W_1) = \sigma_1^2 < \infty$, and $\text{var}(W_2) = \sigma_2^2 < \infty$. Find $E(W_1 - W_2)$ and $\text{var}(W_1 - W_2)$.

Solution:
First, $E(W_1 - W_2) = E(W_1) - E(W_2) = \mu_1 - \mu_2$.
Second, $\text{var}(W_1 - W_2) = E\{[(W_1 - W_2) - E(W_1 - W_2)]^2\}$

$$\text{var}(W_1 - W_2) = E\{[(W_1 - W_2) - (\mu_1 - \mu_2)]^2\}$$
$$\text{var}(W_1 - W_2) = E\{[(W_1 - \mu_1) - (W_2 - \mu_2)]^2\}$$
$$\text{var}(W_1 - W_2) = E\{[(W_1 - \mu_1)^2 + (W_2 - \mu_2)^2 - 2(W_1 - \mu_1)(W_2 - \mu_2)]\}$$

$$\operatorname{var}(W_1 - W_2) = E[(W_1-\mu_1)^2] + E[(W_2-\mu_2)^2] - 2E[(W_1-\mu_1)(W_2-\mu_2)]$$
$$\operatorname{var}(W_1 - W_2) = \operatorname{var}(W_1) + \operatorname{var}(W_2) - 2\operatorname{cov}(W_1, W_2)$$
$$\operatorname{var}(W_1 - W_2) = \sigma_1^2 + \sigma_2^2 - 2 \bullet 0$$
$$\operatorname{var}(W_1 - W_2) = \sigma_1^2 + \sigma_2^2$$

Two Independent Sample Test

Let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* from the random variable *X*, which is $N(\mu_X, \sigma_X^2)$. For example, *X* could be the response of a participant in a clinical trial to a new medicine. Let $B_1, B_2, \ldots, B_m$ be a random sample of size *m* from the random variable *B*, which is $N(\mu_B, \sigma_B^2)$. Continuing the example, B could be the response of a participant in a clinical trial to the best available medicine. The two samples are independent of each other. The context of this discussion is that there is random assignment of the participants in the study to the two groups. This has the effect of roughly balancing the two groups with respect to each and every variable other than the treatment assigned. If there is a significant difference between the two groups, that difference is either a chance event or the causal result of a difference in the treatments.

Back to the probability theory of the analysis, $\overline{X}_n$ is $N(\mu_X, \frac{\sigma_X^2}{n})$, and $\overline{B}_m$ is $N(\mu_B, \frac{\sigma_B^2}{m})$ The two sample averages are independent. We seek to use this data to test $H_0 : E(X-B) = 0$ against the alternative hypothesis $H_1 : E(X-B) \neq 0$ at the α level of significance. Our test statistic is $TS = \overline{X}_n - \overline{B}_m$.

Distribution of the Test Statistic

The distribution of $TS = \overline{X}_n - \overline{B}_m$ is normal. In the case that the random variable *X* is not normal (or the random variable *B* is not normal), then the CLT implies that the distribution of *TS* is approximately normal. From the problem at the start of class, $E(TS) = E(\overline{X}_n - \overline{B}_m) = E(\overline{X}_n) - E(\overline{B}_m) = \mu_X - \mu_B$, and

$$\operatorname{var}(TS) = \operatorname{var}(\overline{X}_n - \overline{B}_m) = \operatorname{var}(\overline{X}_n) + \operatorname{var}(\overline{B}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_B^2}{m}.$$

Null Distribution of the Test Statistic

Since $H_0 : E(X-B) = 0$ which is equivalent to $H_0 : \mu_X = \mu_B$,

$E_0(TS) = E_0(\overline{X}_n - \overline{B}_m) = \mu_X - \mu_B = 0$.

R. A. Fisher argued that the null hypothesis should be that the random variable *X* has exactly the same distribution as the random variable *B*. That is, not only does $\mu_X = \mu_B$ under the null, but also $\sigma_X^2 = \sigma_B^2 = \sigma^2$. In fact, any probabilistic parameter of random variable has the same value for *X* and *B*. Since it is very unusual for a treatment to have a measurable effect, this conception of the null hypothesis is widely used. Using this assumption, $\operatorname{var}_0(TS) = \operatorname{var}_0(\overline{X}_n - \overline{B}_m) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2(\frac{1}{n} + \frac{1}{m})$.

Test when variance known

As in Chapter 5, we test this null hypothesis by putting TS in standard score form:

$$Z = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}.$$

If $\alpha = 0.10$, we reject H_0 when $|Z| \geq 1.645$. If $\alpha = 0.05$, we reject H_0 when $|Z| \geq 1.960$. If $\alpha = 0.01$, we reject H_0 when $|Z| \geq 2.576$.

Test when variances unknown but equal

Just as in Chapter 5, we use an estimate of σ^2 and stretch the critical values an amount determined by the degrees of freedom of our estimate. There are many estimates of σ^2. For example, $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$ and $S_B^2 = \frac{\sum_{i=1}^{n}(B_i - \overline{B}_m)^2}{m-1}$ are unbiased estimates of σ^2, with $n-1$ and $m-1$ degrees of freedom respectively. That is, $E(S_X^2) = E(S_B^2) = \sigma^2$. We use both estimates. Let $S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_B^2}{n+m-2}$.

This estimator has $n+m-2$ degrees of freedom. Then our studentized statistic is

$$T_{n+m-2} = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{S_P^2(\frac{1}{n} + \frac{1}{m})}}.$$

If $\alpha = 0.10$, we reject H_0 when $|T_{n+m-2}| \geq t_{1.645,n+m-2}$. Similarly, if $\alpha = 0.05$, we reject H_0 when $|T_{n+m-2}| \geq t_{1.960,n+m-2}$. If $\alpha = 0.01$, we reject H_0 when $|T_{n+m-2}| \geq t_{2.576,n+m-2}$.

Example Problem: from examination 1, Fall 2016, 2B

A research team took a random sample of 3 observations from a normally distributed random variable *Y* and observed that $\bar{y}_3 = 37.4$ and $s_Y^2 = 42.6$, where $\bar{y}_3$ was the average of the three observations sampled from *Y* and s_Y^2 was the unbiased estimate of $\text{var}(Y)$ (i.e., the divisor in the variance was $n-1$). A second research team took a random sample of 5 observations from a normally distributed random variable X and observed that $\bar{x}_5 = 50.6$ and $s_X^2 = 48.1$, where $\bar{x}_5$ was the average of the five observations sampled from X and s_X^2 was the unbiased estimate of $\text{var}(X)$ (i.e., the divisor in the variance was $n-1$). Test the null hypothesis $H_0 : E(X) = E(Y)$ against the alternative $H_1 : E(X) \neq E(Y)$ at the 0.10, 0.05, and 0.01 levels of significance using the pooled variance t-test. This problem is worth 40 points.

Solution: $s_P^2 = \dfrac{(5-1)48.1+(3-1)42.6}{5+3-2} = 46.27$ on 6 degrees of freedom. The t stretches of 1.645, 1.960, and 2.576 for 6 degrees of freedom are 1.943, 2.447, and 3.707 respectively. The test statistic is $t_{n+m-2} = \dfrac{37.4-50.6-0}{\sqrt{46.27(\frac{1}{5}+\frac{1}{3})}} = -2.637$. Reject the null hypothesis at the 0.10 and 0.05. Accept the null hypothesis at the 0.01 level.

Alternate Problem:

A research team took a random sample of 3 observations from a normally distributed random variable Y and observed that $\bar{y}_3 = 37.4$ and $s_Y^2 = 42.6$, where $\bar{y}_3$ was the average of the three observations sampled from Y and s_Y^2 was the unbiased estimate of $\text{var}(Y)$ (i.e., the divisor in the variance was $n-1$). A second research team took a random sample of 5 observations from a normally distributed random variable X and observed that $\bar{x}_5 = 50.6$ and $s_X^2 = 48.1$, where $\bar{x}_5$ was the average of the five observations sampled from X and s_X^2 was the unbiased estimate of $\text{var}(X)$ (i.e., the divisor in the variance was $n-1$). Find the 99% confidence interval for $E(X)-E(Y)$.

Solution: The template for the 99% confidence interval for $E(X)-E(Y)$ is $\bar{x}_n - \bar{y}_m \pm t_{2.576,n+m-2}\, s_P \sqrt{\dfrac{1}{n}+\dfrac{1}{m}}$. That is, the 99% confidence interval for $E(X)-E(Y)$ is $50.6-37.4 \pm 3.707\sqrt{46.27}\sqrt{\dfrac{1}{5}+\dfrac{1}{3}} = 13.2 \pm 3.707 \bullet 6.802 \bullet 0.7303 = 13.2 \pm 18.4$. That is, the 99% confidence interval for $E(X)-E(Y)$ is the interval between -5.2 and 31.6. Since 0 is in the 99% confidence interval, we should accept $H_0 : E(X) = E(Y)$ against $H_1 : E(X) \neq E(Y)$ at the 0.01 level of significance.

Test when variances unknown and unequal

This procedure is called the unequal variance t-test or unequal variance confidence interval. As before, let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* from the random variable *X*, which is $N(\mu_X, \sigma_X^2)$. Let $B_1, B_2, \ldots, B_m$ be a random sample of size *m* from the random variable *B*, which is $N(\mu_B, \sigma_B^2)$. Here, $\sigma_X^2 \neq \sigma_B^2$. Then,

$$\mathrm{var}(\overline{X}_n - \overline{B}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_B^2}{m}.$$

The standard score form of the test statistic is then

$$Z = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_B^2}{m}}}.$$

The studentized standard score form of test statistic would be

$$T_? = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_B^2}{m}}}.$$

The problem is that the probability theory calculations that are the basis of the degrees of freedom in the pooled variance t-test are not valid. There is, however, a complicated formula (called Satterthwaite's formula) that produces an approximate degrees of freedom. I will not ask you to calculate this by hand in an examination. Fortunately, the statistical programs that you will use will calculate the unequal variance t-test and unequal variance confidence interval for you. Typically, the equal variance t-test and unequal variance t-tests have essentially equal p-values when the assumption of equal variance appears reasonable. When the assumption does not appear reasonable, the one should use the unequal variance calculations. My practice is to report the unequal variance results as calculated by a reputable statistics program.

Type II Error Rate and Power Calculations

The definition of the Type II error rate is $\beta = \Pr_1\{\text{Accept } H_0\}$. The power of a statistical test is defined to be $\text{Power} = 1 - \beta$. This is the probability that the null hypothesis is correctly rejected. A large Type II error rate indicates a study that is "underpowered." The calculation of β is just a normal probability calculation. The specification of the normal distribution is based on the alternative specified in the problem. Typically, the values of the variances are assumed known.

Example Problem: from Chapter 6 Study Guide, Problem 3

In a clinical trial, 50 patients suffering from an illness will be randomly assigned to one of two groups so that 25 receive an experimental treatment and 25 receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable *B* is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed with $\sigma_X = \sigma_B = 500$ under both the null and alternative distributions. The null hypothesis to be tested is that $E(X) - E(B) = 0$ against the alternative that $E(X) - E(B) > 0$ at the 0.01 level of significance. What is the probability of a Type II error for the test of the null hypothesis when $E(X) - E(B) = 500$?

Solution: The standard score form of the TS is $Z = \dfrac{\bar{X}_n - \bar{B}_m - 0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}$, and $H_0 : E(X - B) = 0$ is rejected when $Z \geq 2.326$, remembering that the problem asks for a one-sided test at level of significance 0.01. Using the TS directly, $H_0 : E(X - B) = 0$ is rejected when

$\bar{X}_{25} - \bar{B}_{25} \geq 0 + 2.326\sqrt{\dfrac{500^2}{25} + \dfrac{500^2}{25}} = 0 + 2.326 \bullet 141.42 = 328.95$. For the alternative specified in the problem $\bar{X}_{25} - \bar{B}_{25}$ is $N(500, 141.42^2)$. Then

$$\beta = \Pr_1\{\text{Accept } H_0\} = \Pr_1\{\bar{X}_{25} - \bar{B}_{25} < 328.95\} = \Pr\{\frac{\bar{X}_{25} - \bar{B}_{25} - E_1(\bar{X}_{25} - \bar{B}_{25})}{\sigma_1(\bar{X}_{25} - \bar{B}_{25})} < \frac{328.95 - 500}{141.42}\}.$$

That is, $\beta = \Pr\{Z < \dfrac{328.95 - 500}{141.42} = -1.210\} = \Phi(-1.210) = 0.113$.

Sample size for two sample test:

The bad news is that the mathematics of sample size calculations is relatively complex. The good news is that one can solve a wide range of sample size problems once one knows how to solve this one. The argument is essentially the same.

Problem 6 in Chapter 6 Study Guide

> In a clinical trial, *2J* patients suffering from an illness will be randomly assigned to one of two groups so that *J* will receive an experimental treatment and *J* will receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable

B is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed. The null hypothesis to be tested is that $E(X)-E(B)=0$ against the alternative that $E(X)-E(B)>0$ at the $\alpha,\ \alpha\le 0.5,$ level of significance. When the null hypothesis is true, $\mathrm{var}(X)=\mathrm{var}(B)=\sigma_0^2$. When the alternative hypothesis is true, $\mathrm{var}(B)=\sigma_0^2$, but $\mathrm{var}(X)=\sigma_1^2>\sigma_0^2$. What is the number *J* in each group that would have to be taken so that the probability of a Type II error for the test of the null hypothesis specified in the common section is $\beta,\ \beta\le 0.5,$ when $E(X)-E(B)=\Delta>0$?

Solution: The test statistic is $TS=\overline{X}_J-\overline{B}_J$, and *TS* is $N(E(X)-E(B),\dfrac{\mathrm{var}(X)}{J}+\dfrac{\mathrm{var}(B)}{J})$. The null distribution of *TS* is then $N(0,\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_0^2}{J})$. Hence, we reject $H_0:E(X)-E(B)=0$ against $H_1:E(X)-E(B)>0$ at the α level of significance when $TS\ge 0+|z_\alpha|\sqrt{\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_0^2}{J}}$. When $E(X)-E(B)=\Delta>0$ and $\mathrm{var}(X)=\sigma_1^2>\sigma_0^2.\ \mathrm{var}(B)=\sigma_0^2,$ the (alternative) distribution of *TS* is then $N(\Delta,\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_1^2}{J})$. Then, the probability of a Type II error is $\beta=\Pr_1\{\text{Accept } H_0\}=\Pr_1\{TS<0+|z_\alpha|\sqrt{\dfrac{2\sigma_0^2}{J}}\}$. That is,

$$\beta=\Pr_1\{TS<0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}\}=\Pr\{Z=\frac{TS-\Delta}{\sigma_1(\overline{X}_J-\overline{B}_J)}<\frac{0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}}\}.$$ Since

$\beta=\Pr_1\{\text{Accept } H_0\}\le 0.5$, it is true that $\beta=\Pr\{Z<-|z_\beta|\}$. We now have two equations:

$$\beta=\Pr\{Z<\frac{0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}}\},\text{ and}$$

$$\beta=\Pr\{Z<-|z_\beta|\}.$$

The problem is to choose *J* so that the probability of a Type II error is a specified value. That is, we should choose *J* so that the right-hand sides of the two equations are equal: $\dfrac{0+|z_\alpha|\sqrt{\dfrac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_1^2}{J}}}=-|z_\beta|$.

We have to solve for *J* in the equation above. This reduces to: $0+|z_\alpha|\sqrt{\dfrac{2\sigma_0^2}{J}}-\Delta=-|z_\beta|\sqrt{\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_1^2}{J}}$. That is, $|z_\alpha|\sqrt{\dfrac{2\sigma_0^2}{J}}+|z_\beta|\sqrt{\dfrac{\sigma_0^2}{J}+\dfrac{\sigma_1^2}{J}}=\Delta$. Next, solve for *J* to get $\sqrt{J}=\dfrac{|z_\alpha|\sqrt{2\sigma_0^2}+|z_\beta|\sqrt{\sigma_0^2+\sigma_1^2}}{\Delta}$. Since *J* has to be an integer, we increase *J* to the next integer value.

The scenario can be realistic. Suppose that there are two ratio scale random variables. The one with the greater mean typically has a greater variance. For example, the Poisson and the chi-square distributions have this property. This greater variance requires a somewhat greater sample size in study design.

This calculation assumes that there is no attrition of subjects. Typically, study attrition is large. An attrition rate less than 15% at a follow-up three or more years later is a very low attrition rate. Accounting for attrition is its own modeling effort, which can be very difficult to do well.

Problem 4 in Chapter 6 Study Guide

In a clinical trial, *2J* patients suffering from an illness will be randomly assigned to one of two groups so that *J* will receive an experimental treatment and *J* will receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable *B* is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed and have $\sigma_X=\sigma_B=500$ under both the null and alternative distributions. The null hypothesis to be tested is that $E(X)-E(B)=0$ against the alternative that $E(X)-E(B)>0$ at the 0.005 level of significance. What is the number *J* in each group that would have to be taken so that the probability of a Type II error for the test of the null hypothesis specified in the common section is 0.01 when $E(X)-E(B)=250$?

Solution: For this specification, $\alpha = .005$, so that $|z_\alpha| = 2.576$. Also, $\beta = .01$, so that $|z_\beta| = 2.326$. With regard to variances, $\sigma_0^2 = \sigma_1^2 = 500^2$. Finally, $\Delta = 250$. Then, the design equation is $\sqrt{J} = \frac{2.576\sqrt{2 \bullet 500^2} + 2.326\sqrt{2 \bullet 500^2}}{250} = \frac{3466.24}{250} = 13.865 = \sqrt{192.24}$.
That is, there should be at least 193 in each group.

The magnitude of the difference in the expected values is 250, which is half of the assumed standard deviation of each group. This is called a one-half standard deviation effect. To detect a difference equal to 1/2 of the standard deviation, researchers need about 200 in each group. This is 400 total observations. One needs about 50 observations per group to detect a one standard deviation effect. That is, fewer observations are needed to detect a larger effect size.

Chapter 7
Inferences about Population Variances

Probability Theory Facts

1. Let Z be $N(0,1)$. Then Z^2 has the (central) chi-squared distribution with 1 degree of freedom. This is denoted χ_1^2.

2 Let $Z_1, Z_2, \ldots, Z_n$ be $NID(0,1)$. Then $S_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$ follows the (central) chi-square distribution with *n* degrees of freedom, denoted χ_n^2. The expected value of a χ_n^2 is *n*: $E(S_n) = E(Z_1^2) + E(Z_2^2) + \cdots + E(Z_n^2)$. Since $\text{var}(Z) = 1 = E(Z^2) - [E(Z)]^2 = 1$, then $E(Z^2) - [0]^2 = 1$. Using this in $E(S_n) = E(Z_1^2) + E(Z_2^2) + \cdots + E(Z_n^2) = n$. Further, the variance of a chi-square distribution with *n* degrees of freedom is *2n*: $\text{var}(S_n) = 2n$

3 Let Y be $N(\mu_Y, \sigma_Y^2)$. Then, $\frac{Y - \mu_Y}{\sigma_Y} = Z$ is $N(0,1)$. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from Y, which is $N(\mu_Y, \sigma_Y^2)$. Then $\sum_{i=1}^{n} (\frac{Y_i - \mu_Y}{\sigma_Y})^2$ is χ_n^2. After factoring out σ_Y^2, $\frac{\sum_{i=1}^{n} (Y_i - \mu_Y)^2}{\sigma_Y^2}$ is also χ_n^2.

Since μ_Y is not known in applications, it must be estimated. An important property of a sample from a normal distribution is that $\dfrac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}{\sigma_Y^2}$ is distributed as χ_{n-1}^2. That is, using the sample mean has reduced the degrees of freedom by one. From AMS 310, the unbiased estimator of the sample variance is $S^2 = \dfrac{\sum(Y_i - \overline{Y}_n)^2}{n-1}$. Since $(n-1)S^2 = (n-1)\dfrac{\sum(Y_i - \overline{Y}_n)^2}{n-1} = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$, $\dfrac{(n-1)S^2}{\sigma_Y^2}$ has a central chi-squared distribution with $n-1$ degrees of freedom when $Y_1, Y_2, \ldots, Y_n$ is a sample of size n from a $N(\mu_Y, \sigma_Y^2)$ distribution. This is our first important use of the chi-squared distribution. The tests in this chapter are usually one-sided.

Problem 1 from Chapter 7 Study Guide

A research team took a sample of 8 observations from the random variable *Y*, which had a normal distribution $N(\mu, \sigma^2)$. They observed $\bar{y}_8 = 43.2$, where $\bar{y}_8$ is the average of the eight sampled observations and $s^2 = 517.5$ is the observed value of the unbiased estimate of σ^2, based on the sample values. Test the null hypothesis that $H_0 : \sigma^2 = 400$ against the alternative $H_1 : \sigma^2 > 400$ at the 0.10, 0.05, and 0.01 levels of significance.

Solution: The test statistic is $TS = \dfrac{(n-1)S^2}{\sigma_Y^2}$, which has a central χ_{n-1}^2, where $n-1 = 8-1 = 7$. Since $H_0 : \sigma^2 = 400$, the null distribution of $TS = \dfrac{(n-1)S^2}{\sigma_Y^2} = \dfrac{(n-1)S^2}{400}$ is χ_7^2. The problem specifies a right-sided test with levels of significance 0.10, 0.05, and 0.01. From Table 7, the right-sided critical values are 12.02 (for the 0.10 level), 14.07 (for 0.05), and 18.48 (for 0.01). The value of the chi-squared test statistic is $ts == \dfrac{(8-1) \bullet 517.5}{400} = 9.056$. Since this is less than each of the critical values, we accept the null hypothesis at the 0.10, 0.05, and 0.01 levels. The value of the sample mean is not relevant. A common mistake is for a student to take the sample mean as a cue and answer with a one-sample t-test. This is not correct.

As usual, a confidence interval may be more informative than a statistical test. The next problem is an example of finding a confidence interval for a variance.

Problem 2 from Chapter 7 Study Guide

A research team took a sample of 7 observations from the random variable Y, which had a normal distribution $N(\mu,\sigma^2)$. They observed $\bar{y}_7 = 93.4$, where $\bar{y}_7$ was the average of the sampled observations, and $s^2 = 47.5$ was the observed value of the unbiased estimate of σ^2, based on the sample values. Find the 99% confidence interval for σ^2.

Solution: Since the estimated variance has 6 degrees of freedom,

$$\Pr\{0.6757 < \frac{\sum_{i=1}^{7}(Y_i - \bar{Y}_7)^2}{\sigma^2} < 18.55\} = \Pr\{0.6757 < \frac{6S^2}{\sigma^2} < 18.55\} = 0.99$$. Then

$$\Pr\{0.6757 < \frac{6S^2}{\sigma^2} < 18.55\} = \Pr\{\frac{1}{18.55} < \frac{\sigma^2}{6S^2} < \frac{1}{0.6757}\} = \Pr\{\frac{6S^2}{18.55} < \sigma^2 < \frac{6S^2}{0.6757}\}.$$

The interval from $\frac{6S^2}{18.55}$ to $\frac{6S^2}{0.6757}$ is then the basis of a confidence interval for σ^2. In this problem, the left end of the confidence interval is $\frac{6s^2}{18.55} = 0.3235s^2 = 0.3235 \bullet 47.5 = 15.36$, and the right end is $\frac{6s^2}{0.6757} = 8.880s^2 = 8.880 \bullet 47.5 = 421.785$. The confidence interval extends from a factor of about 3 less than $s^2 = 47.5$ to a factor of about 9 greater than $s^2 = 47.5$. Again, the sample mean is not needed to answer the question.

When you work these problems, examine your answer and notice that the confidence interval for σ^2 is very wide. Specifically examine the ratio of the upper limit to the lower limit, here almost 28. It is remarkable that the t distribution stretches are as small as they are. One gets percentiles of the chi-squared distribution from Table 7. The Excel spreadsheet and all statistical packages have the percentiles available as well.