Chapter 5: Introduction to Inference: Estimation and Prediction

The PICTURE

In Chapters 1 and 2, we learned some basic methods for analyzing the pattern of variation in a set of data. In Chapter 3, we learned that in order to take statistics to the next level, we have to design studies to answer specific questions. We also learned something about designed studies. In Chapter 4, we learned about statistical models and the mathematical language in which they are written: probability.

We are now ready to move to the next level of statistical analysis: **statistical inference**. Statistical inference uses a sample from a population to draw conclusions about the entire population. The material of Chapter 3 enables us to obtain the sample in a statistically valid way. The models and probabilistic concepts of Chapter 4 enable us to obtain valid inference and to quantify the precision of the results obtained.

Preview:

- The models: C+E and binomial
- Types of inference: estimation, prediction, tolerance interval
- Estimation basics:
  - o Estimator or estimate?
  - o Sampling distributions
  - o Confidence intervals
- Estimation for the one population C+E model
- Prediction for the one population C+E model
- Estimation for the one population binomial model
- Determination of sample size
- Estimation for the two population C+E model
- Estimation for the two population binomial model
- Normal theory tolerance intervals

Statistical Inference:

Use of a subset of a population (the sample) to draw conclusions about the entire population.

The validity of inference is related to the way the data are obtained, and to the stationarity of the process producing the data.

For valid inference the units on which observations are made must be obtained using a **probability sample**. The simplest probability sample is a **simple random sample (SRS)**.

The Models

We will study

- The C+E model

  $$Y = \mu + \epsilon,$$

  where μ is a **model parameter** representing the center of the population, and ϵ is a random error term (hence the name C+E).
  Often, we assume that $\epsilon \sim N(0, \sigma^2)$, which implies $Y \sim N(\mu, \sigma^2)$.
- The binomial model

The Data

Before they are obtained, the data are represented as independent random variables, $Y_1, Y_2, \ldots, Y_n$. After they are obtained, the resulting values are denoted $y_1, y_2, \ldots, y_n$.

For the C+E model, the data are considered a random sample from a population described by the C+E model (e.g, $N(\mu, \sigma^2)$).

The binomial model represents data from a population in which the sampling units are observed as either having or not having a certain characteristic. The proportion of the population having the characteristic is p.

n sampling units are drawn randomly from the population. Y_i equals 1 if sampling unit i has the characteristic, and 0 otherwise. Therefore, $Y_1, Y_2, \ldots, Y_n$ are independent $b(1, p)$ (also known as Bernoulli(p)) random variables. For the purpose of inference about p, statistical theory says that we need only consider $Y = \sum_{i=1}^{n} Y_i$, the total number of sampling units in the sample that have the characteristic. In Chapter 4 we learned that $Y \sim b(n, p)$.

Types of Inference

- Estimation of model parameters
- Prediction of a future observation
- Tolerance interval

Point Estimation for μ in the C+E Model

- **Least absolute errors** finds m to minimize

$$\mathrm{SAE}(m) = \sum_{i=1}^{n} |y_i - m|.$$

  For the C+E model, the least absolute errors estimator is the sample median, Q_2.

- **Least squares** finds m to minimize

$$\mathrm{SSE}(m) = \sum_{i=1}^{n} (y_i - m)^2.$$

  For the C+E model, the least squares estimator is the sample mean, $\overline{Y}$.

TYU 22

Estimator or Estimate?

- The Randomness in a set of data from a designed study is in the production of the data: measuring, sampling, treatment assignment, etc.
- An **estimator** is a **rule** for computing a quantity **from a sample** that is to be used to estimate a model parameter.
- An **estimate** is the **value** that rule gives **when the data are taken**.

Estimation for the C+E Model: Sampling Distributions

The distribution model of an estimator is called its **sampling distribution**. For example, in the C+E model, the least squares estimator $\overline{Y}$, has a $N(\mu, \sigma^2/n)$ distribution (its sampling distribution):

- Exactly, if $\epsilon \sim N(0, \sigma^2)$
- Approximately, if n is large enough. The CLT guarantees it!

One further bit of terminology: The standard deviation of the sampling distribution of an estimator is called the **standard error** of the estimator. So, the standard error of $\overline{Y}$ is $\sigma/\sqrt{n}$.

Confidence Intervals

A level L **confidence interval** for a parameter θ is an interval $(\hat{\theta}_1,\ \hat{\theta}_2)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimators having the property that

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = L.$$

Estimation for the C+E Model: Confidence Interval for μ: Known Variance

Suppose we know σ^2. Then if $\overline{Y}$ can be assumed to have a $N(\mu, \sigma^2/n)$ sampling distribution, we know that

$$Z = \frac{(\overline{Y} - \mu)}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma}$$

has a $N(0,1)$ distribution.

Let z_δ denote the δ quantile of the standard normal distribution: i.e., if $Z \sim N(0,1)$, then $P(Z \leq z_\delta) = \delta$. Then

$$\begin{aligned} L &= P\left(z_{(1-L)/2} < \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} < z_{(1+L)/2}\right) \\ &= P\left(\overline{Y} - \frac{\sigma}{\sqrt{n}} z_{(1+L)/2} < \mu < \overline{Y} - \frac{\sigma}{\sqrt{n}} z_{(1-L)/2}\right). \end{aligned}$$

Noting that

$$z_{\frac{1-L}{2}} = -z_{\frac{1+L}{2}},$$

we obtain the formula for a level L confidence interval for μ:

$$\left(\overline{Y} - \frac{\sigma}{\sqrt{n}} z_{\frac{1+L}{2}}, \overline{Y} + \frac{\sigma}{\sqrt{n}} z_{\frac{1+L}{2}}\right).$$

Denoting the standard error of $\overline{Y}$, $\sigma/\sqrt{n}$, by $\sigma(\overline{Y})$, we have the formula

$$\left(\overline{Y} - \sigma(\overline{Y}) z_{\frac{1+L}{2}}, \overline{Y} + \sigma(\overline{Y}) z_{\frac{1+L}{2}}\right).$$

Example 1:

A computer scientist is investigating the usefulness of a design language in improving programming tasks. Twelve expert programmers are asked to code a standard function in the language, and the times taken to complete the task (in minutes) are recorded. The data are:
17 16 21 14 18 24 16 14 21 23 13 18
We will assume these data were generated by the C+E model:

$$Y = \mu + \epsilon.$$

We first have a look at the data and check the assumption of normality.

The point estimate of μ is $\overline{y} = 17.9167$.
Suppose we know $\sigma = 3.6296$. Then

$$\sigma(\overline{Y}) = \frac{\sigma}{\sqrt{n}} = \frac{3.6296}{\sqrt{12}} = 1.0478,$$

and a 95% confidence interval for μ is

$$\big(\overline{Y} - \sigma(\overline{Y})z_{0.975}, \overline{Y} + \sigma(\overline{Y})z_{0.975}\big)$$

$$= (17.9167 - (1.0478)(1.96), 17.9167 + (1.0478)(1.96))$$

$$= (15.8630, 19.9704).$$

Based on these data, we estimate that μ lies in the interval (15.8630,19.9704).

The Interpretation of Confidence Level

The confidence level, L, of a level L confidence interval for a parameter θ is interpreted as follows: Consider all possible samples that can be taken from the population described by θ and for each sample imagine constructing a level L confidence interval for θ. Then a proportion L of all the constructed intervals will really contain θ.

Example 1, Continued:

Recall that based on a random sample of 12 programming times, we computed a 95% confidence interval for μ, the mean programming time for all programmers in the population from which the sample was drawn. (15.8630,19.9704). We are 95% confident in our conclusion, meaning that in repeated sampling, 95% of all intervals computed in this way will contain the true value of μ.

Demo Time!

Recap:

Recall that we developed a confidence interval for the population mean μ under the assumption that the population standard deviation σ was known, by using the fact that

$$\frac{\overline{Y} - \mu}{\sigma(\overline{Y})} \sim N(0, 1),$$

at least approximately, where $\overline{Y}$ is the sample mean and $\sigma(\overline{Y}) = \sigma/\sqrt{n}$ is its standard error.
The level L confidence interval we developed was

$$\left(\overline{Y} - \sigma(\overline{Y}) z_{\frac{1+L}{2}}, \overline{Y} + \sigma(\overline{Y}) z_{\frac{1+L}{2}}\right),$$

where $z_{\frac{1+L}{2}}$ is the $(1 + L)/2$ quantile of the $N(0, 1)$ distribution.

Estimation for the C+E Model: Confidence Interval for μ: Unkown Variance

If σ is unknown, estimate it using the sample standard deviation, S. This means that instead of computing the exact standard error of $\overline{Y}$, we use the estimated standard error,

$$\hat{\sigma}(\overline{Y}) = \frac{S}{\sqrt{n}}.$$

However, the resulting standardized estimator,

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}(\overline{Y})},$$

now has a t_{n-1}, rather than a $N(0,1)$, distribution. The result is that a level L confidence interval for μ is given by

$$\left(\overline{Y} - \hat{\sigma}(\overline{Y}) t_{n-1,\frac{1+L}{2}}, \overline{Y} + \hat{\sigma}(\overline{Y}) t_{n-1,\frac{1+L}{2}}\right).$$

Example 1, Continued:

Recall again the programming time example. In reality, we don't know σ, but we can estimate it using the sample standard deviation, S.

For these data, $n = 12$ and $s = 3.6296$, which means that $\hat{\sigma}(\overline{Y}) = \frac{3.6296}{\sqrt{12}} = 1.0478$. In addition, $t_{n-1,\frac{1+L}{2}} = t_{11,0.975} = 2.2010$, so a level 0.95 confidence interval for μ is

$$= (17.9167 - (1.0478)(2.2010), 17.9167 + (1.0478)(2.2010))$$

$$= (15.6105, 20.2228).$$

This interval is slightly wider than the previous interval, because it must account for the additional uncertainty in estimating σ. This is reflected in the larger value of $t_{11,0.975} = 2.2010$ compared with $z_{0.975} = 1.96$.

TYU 23

Prediction for the C+E Model

The problem is to use the sample to predict a new observation (i.e. one that is not in the sample) from the C+E model. This is a very different problem than estimating a model parameter, such as μ.

To see what is involved, call the new observation Y_{new}. We know that $Y_{new} = \mu + \epsilon_{new}$. Let $\hat{Y}_{new}$ denote a predictor of Y_{new}. Suppose for now that we know μ. Then it can be shown that the "best" predictor is $\hat{Y}_{new} = \mu$. However, even using this knowledge, we will still have prediction error:

$$Y_{new} - \hat{Y}_{new} = Y_{new} - \mu = (\mu + \epsilon_{new}) - \mu = \epsilon_{new}.$$

The variance of prediction, $\sigma^2(Y_{new} - \hat{Y}_{new})$, is therefore σ^2, the variance of the model's error distribution.

We won't know μ, however, so we use $\overline{Y}$ from the sample to estimate it, giving the predictor $\hat{Y}_{new} = \overline{Y}$. The prediction error is then

$$Y_{new} - \hat{Y}_{new} = (\mu + \epsilon_{new}) - \hat{Y}_{new} = (\mu - \hat{Y}_{new}) + \epsilon_{new}$$

$$= (\mu - \overline{Y}) + \epsilon_{new}$$

$\mu - \overline{Y}$ is the error due to using $\overline{Y}$ to estimate μ. Its variance, as we have already seen, is σ^2/n. ϵ_{new} is the random error inherent in Y_{new}. Its variance is σ^2. Since these terms are independent, the variance of their sum is the sum of their variances (see text, ch. 4):

$$\begin{aligned}
\sigma^2(Y_{new} - \hat{Y}_{new}) &= \sigma^2(\mu - \hat{Y}_{new}) + \sigma^2(\epsilon_{new}) \\
&= \frac{\sigma^2}{n} + \sigma^2 \\
&= \sigma^2\left[1 + \frac{1}{n}\right].
\end{aligned}$$

In most applications σ will not be known, so we estimate it with the sample standard deviation S, giving the estimated standard error of prediction

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = S\sqrt{1 + \frac{1}{n}}.$$

A level L prediction interval for a new observation is then

$$\hat{Y}_{new} \pm \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-1,\frac{1+L}{2}}.$$

Example 1, Continued:

We return to the programming time example. Recall that for these data, $\overline{y} = 17.9167$, so that the predicted value is $\hat{y}_{new} = \overline{y} = 17.9167$. Also, $n = 12$ and $s = 3.6296$, which means that

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = 3.6296\sqrt{1 + \frac{1}{12}} = 3.77781.$$

In addition, $t_{n-1,\frac{1+L}{2}} = t_{11,0.975} = 2.2010$, so a level 0.95 prediction interval for the diameter of a new piece is:

$$(17.9167 - (3.77781)(2.2010), 17.9167 + (3.77781)(2.2010))$$

$$= (9.60175, 26.2317).$$

Notice how much wider this is than the confidence interval we obtained for μ: (15.6105,20.2228).

Interpretation of the Confidence Level for a Prediction Interval

The interpretation of the level of confidence L for a prediction interval is similar to that for a confidence interval: Consider all possible samples that can be taken from the population, and for each sample imagine constructing a level L prediction interval for a new observation. Also, for each sample draw a 'new' observation at random from among the remaining population values not in the sample. Then a proportion L of all the constructed intervals will really contain their 'new' observation.

Demo Time!

What's the IDEA?

When deciding between a confidence or prediction interval, ask whether you are estimating a model parameter or predicting a new observation:

A Confidence Interval is a range of plausible values for a model parameter (such as the mean, or as we will see later, a population proportion). Since the population parameter is fixed, the only variation involved is the variation in the data used to construct the interval.

A Prediction Interval is a range of plausible values for a new observation (i.e., one not in the sample) selected at random from the population. The prediction interval is wider than the corresponding confidence interval since, in addition to the variation in the sample, it must account for the variation involved in obtaining the new observation.

TYU 24

Point Estimation of a Population Proportion Example 2:

We'll once again consider the grinding example from Chapter 4. In order to determine the correctness of their suspicion that parts tended to be ground larger than spec, management had 150 parts sampled at random. If Y is the number of the 150 parts that exceed spec, then $Y \sim b(150, p)$, where p is the population proportion of parts that exceed spec.

A point estimate of p is then the sample proportion of parts exceeding spec: $\hat{p} = Y/150$.

Recall that 93 of the 150 parts in the sample exceeded spec, giving a point estimate $\hat{p} = 93/150 = 0.62$. So we estimate that 62% of all parts exceed spec.

In general, if we sample n items from a population in which a proportion p have a certain characteristic of interest, and if Y is the number in the sample that have the characteristic, then $Y \sim b(n, p)$. The point estimator of p is then the sample proportion, $\hat{p} = Y/n$.

Interval Estimation of a Population Proportion

NOTE: Recent research on confidence intervals for a population proportion p shows that an interval known as the **score interval** gives superior performance to both the exact and large sample intervals described in the text. Further, the same interval works for both large and small samples.

The score interval has a rather complicated formula. The SAS macro *bici* computes the score interval (along with the exact, large-sample (Wald) and yet another interval known as the bootstrap interval (presented in Chapter 11 of the text)).

However, there is an **approximate score interval** which performs nearly as well as the score interval, and which has a simple formula, and is therefore suitable for hand calculation.
The approximate score interval is just the large sample (or Wald) interval (pp. 255-256 of the text) with the sample proportion $\hat{p} = Y/n$ replaced by an adjusted value $\tilde{p}$. The adjustment moves $\hat{p}$ closer to 0.5.
We present this approximate score interval below.

I will require that you use either the score interval or the approximate score interval in homework and tests. More detailed information on the score interval may be found on the course web site under *Other Resources*. See the links for *Revised Confidence Interval Guide*, *The Score Interval for Population Proportions*, and *Analysis Guide for Statistical Intervals*.

The Approximate Score Interval

The level L approximate score interval for p is computed as follows:

1. Compute the adjusted estimate of p:

$$\tilde{p} = \frac{y + 0.5z^2_{(1+L)/2}}{n + z^2_{(1+L)/2}},$$

   where y is the observed number of successes in the sample.

2. The approximate score interval is obtained by substituting $\tilde{p}$ for $\hat{p}$ in the large sample confidence interval formula (found in the text, p. 256):

$$\tilde{p} \pm \hat{\sigma}(\tilde{p}) z_{(1+L)/2},$$

   where

$$\hat{\sigma}(\tilde{p}) = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}.$$

Example 2, Continued:

We'll once again consider the grinding example from Chapter 4. Recall that 150 parts were sampled at random and that 93 had diameters greater than the specification diameter.

We will use these data to obtain a level 0.99 approximate score confidence interval for *p*, the true population proportion of parts with diameters greater than spec.

The approximate score interval is computed as follows: Since $L = 0.99$, $z_{(1+L)/2} = z_{0.995} = 2.5758$. Using this in the formula, we obtain

$$\tilde{p} = \frac{93 + (0.5)(2.5758^2)}{150 + 2.5758^2} = 0.6149,$$

so the interval is

$$0.6149 \pm 2.5758\sqrt{\frac{0.6149(1 - 0.6149)}{150}}$$

$$= (0.51, 0.72)$$

Since the interval contains only values exceeding 0.5, we can conclude with 99% confidence that more than half the population diameters exceed spec.

TYU 25

What's the IDEA?

All confidence intervals you will see in this chapter have the same form:
ESTIMATOR ± MULTIPLIER ×
ESTIMATED STANDARD ERROR OF ESTIMATOR
The multiplier is based on the sampling distribution of the estimator and the specified confidence level.

Determination of Sample Size

One consideration in designing an experiment or sampling study is the **precision** desired in estimators or predictors. Precision of an estimator is a measure of how variable that estimator is. Another equivalent way of expressing precision is the width of a level L confidence interval. For a given population, precision is a function of the size of the sample: the larger the sample, the greater the precision.

Suppose it is desired to estimate a population proportion p to within d units with confidence level at least L. Assume also that we will be using an approximate score interval. The requirement is that one half the length of the confidence interval equal d, or

$$z_{\frac{1+L}{2}}\sqrt{\tilde{p}(1-\tilde{p})/n} = d$$

Solving this equation for n gives the required sample size as

$$n = (\tilde{p}(1-\tilde{p}) \cdot z^2_{\frac{1+L}{2}})/d^2$$

We can get an estimate of $\tilde{p}$ from a pilot experiment or study: use the sample proportion $\hat{p}$. Or, since $\tilde{p}(1-\tilde{p}) \leq .25$, we can use .25 in place of $\tilde{p}(1-\tilde{p})$ in the formula.

There is an analogous formula when a simple random sample will be used and it is desired to estimate a population mean μ to within d units with confidence level at least L. If we assume the population is normal, or if we have a large enough sample size (so the normal approximation can be used in computing the confidence interval), the required sample size is

$$n = (\sigma^2 \cdot z^2_{\frac{1+L}{2}})/d^2.$$

Again, this supposes we know σ^2. If we don't, we can get an estimate from a pilot experiment or study.

Example 3:

Suppose we want to use a level 0.90 approximate score interval to estimate to within 0.05 the proportion of voters who support universal health insurance. A pilot survey of 100 voters finds 35 who support universal health insurance. For simplicity, we use the sample proportion $\hat{p} = 0.35$ to estimate $\tilde{p}$. Then, since $d = 0.05$, $L = 0.90$, and $z_{\frac{1+L}{2}} = z_{0.95} = 1.645$, the necessary sample size is

$$n = (0.35(1 - 0.35) \cdot 1.645^2)/0.05^2 = 246.24,$$

so we take $n = 247$.

TYU 26

The Two Population C+E Model

We assume that there are n_1 measurements from population 1 generated by the C+E model

$$Y_{1,i} = \mu_1 + \epsilon_{1,i},\ i = 1, \ldots, n_1,$$

and n_2 measurements from population 2 generated by the C+E model

$$Y_{2,i} = \mu_2 + \epsilon_{2,i},\ i = 1, \ldots, n_2.$$

We want to compare μ_1 and μ_2.

Estimation for Paired Comparisons

Sometimes each observation from population 1 is paired with another observation from population 2. For example, we may want to assess the gain in student knowledge after a course is taken. To do so, we may give each student a pre-course and post-course test. The two populations are then the pre-course and post-course scores, and they are paired by student.

In this case $n_1 = n_2$ and by looking at the pairwise differences, $D_i = Y_{1,i} - Y_{2,i}$, we transform the two population problem to a one population problem for C+E model $D = \mu_D + \epsilon_D$, where $\mu_D = \mu_1 - \mu_2$ and $\epsilon_D = \epsilon_1 - \epsilon_2$. Therefore, a confidence interval for $\mu_1 - \mu_2$ is obtained by constructing a one sample confidence interval for μ_D.

Example 4:

Recall the data set from Example 1, which consisted of programming times for 12 programmers using a particular design language. These data were part of a more extensive study which also obtained the times it took the same 12 programmers to program the same function using a second design language. The researcher wanted to compare the mean programming times in the two design languages. To do so, he computed D, the difference between the programmer's programming time using language 1 and that using language 2. Assuming that these differences follow a C+E model, he constructed a level 0.95 confidence interval for the mean difference in programming time, μ_D. The data (found in SASDATA.PROGRAM_TIMES) are:

	LANGUAGE		
PROGRAMMER	1	2	DIFF
1	17	18	−1
2	16	14	2
3	21	19	2
4	14	11	3
5	18	23	−5
6	24	21	3
7	16	10	6
8	14	13	1
9	21	19	2
10	23	24	−1
11	13	15	−2
12	18	20	−2

An inspection of the differences shows no evidence of nonnormality or outliers. For these data, $\overline{d} = 0.6667$, $s_d = 2.9644$ and $t_{11,0.975} = 2.201$. Then $\hat{\sigma}(\overline{D}) = 2.9644/\sqrt{12} = 0.8558$, so the desired interval is

$$0.6667 \pm (0.8558)(2.201) = (-1.2168, 2.5502).$$

Based on this, we estimate that the mean time to program the function in question using design language 1 is between 2.5502 minutes greater than and 1.2168 minutes less than it takes using design language 2. In particular, since the interval contains 0, we are unable to conclude that there is a difference in mean programming time.

TYU 27

Estimation for Independent Populations

Let $\overline{Y}_1$ and $\overline{Y}_2$ denote the sample means from populations 1 and 2, S_1^2 and S_2^2 the sample variances. The point estimator of $\mu_1 - \mu_2$ is $\overline{Y}_1 - \overline{Y}_2$.

Equal Variances

If the population variances are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then we estimate σ^2 by the pooled variance estimator

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

The estimated standard error of $\overline{Y}_1 - \overline{Y}_2$ is then given by

$$\hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2) = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

$$t^{(p)} = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2)}$$

has a $t_{n_1+n_2-2}$ distribution. This leads to a level L pooled variance confidence interval for $\mu_1 - \mu_2$:

$$\overline{Y}_1 - \overline{Y}_2 \pm \hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2) t_{n_1+n_2-2, \frac{1+L}{2}}$$

Unequal Variances

If $\sigma_1^2 \neq \sigma_2^2$, an approximate level L confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \pm \hat{\sigma}(\overline{Y}_1 - \overline{Y}_2) t_{\nu, \frac{1+L}{2}},$$

where ν is the largest integer less than or equal to

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}},$$

and

$$\hat{\sigma}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Example 5:

A company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared. A summary of the data shows the following (units are hours):

Manufacturer	n	$\overline{y}$	s
1	10	118.4	26.9
2	13	134.9	18.4

The investigators generated histograms and normal quantile plots of the two data sets and found no evidence of nonnormality or outliers. The point estimate of $\mu_1 - \mu_2$ is $\overline{y}_1 - \overline{y}_2 = 118.4 - 134.9 = -16.5$. They decided to obtain a level 0.90 confidence interval to compare the mean lifetimes of blades from the two manufacturers.

- Pooled variance interval The pooled variance estimate is

$$s_p^2 = \frac{(10-1)(26.9)^2 + (13-1)(18.4)^2}{10+13-2} = 503.6.$$

This gives the estimate of the standard error of $\overline{Y}_1 - \overline{Y}_2$ as

$$\hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2) = \sqrt{503.6\left(\frac{1}{10} + \frac{1}{13}\right)} = 9.44.$$

Finally, $t_{21,0.95} = 1.7207$. So a level 0.90 confidence interval for $\mu_1 - \mu_2$ is

$$(-16.5 - (9.44)(1.7207),\ -16.5 + (9.44)(1.7207))$$

$$= (-32.7, -0.3).$$

- Separate variance interval The estimate of the standard error of $\overline{Y}_1 - \overline{Y}_2$ is

$$\hat{\sigma}(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}} = 9.92.$$

The degrees of freedom ν is computed as the greatest integer less than or equal to

$$\frac{\left(\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}\right)^2}{\frac{\left(\frac{(26.9)^2}{10}\right)^2}{10-1} + \frac{\left(\frac{(18.4)^2}{13}\right)^2}{13-1}} = 15.17,$$

so $\nu = 15$. Finally, $t_{15,0.95} = 1.7530$. So a level 0.90 confidence interval for $\mu_1 - \mu_2$ is

$$(-16.5 - (9.92)(1.753),\ -16.5 + (9.92)(1.753))$$

$$= (-33.9, 0.89).$$

There seems to be a problem here. The pooled variance interval, $(-32.7, -0.3)$, does not contain 0, and so suggests that $\mu_1 \neq \mu_2$. On the other hand, the separate variance interval, $(-33.9, 0.89)$, contains 0, and so suggests we cannot conclude that $\mu_1 \neq \mu_2$.
What to do?
Since both intervals are similar and have upper limits very close to 0, I would suggest taking more data to resolve the ambiguity.

Recap: Estimation of Difference in Means of Two Independent Populations

We assume the measurements from the two populations are generated by the C+E models

$$Y_{1,i} = \mu_1 + \epsilon_{1,i}, \; i = 1, \ldots, n_1,$$

$$Y_{2,i} = \mu_2 + \epsilon_{2,i}, \; i = 1, \ldots, n_2.$$

Let $\overline{Y}_1$ and $\overline{Y}_2$ denote the sample means from populations 1 and 2, S_1^2 and S_2^2 the sample variances. The point estimator of $\mu_1 - \mu_2$ is $\overline{Y}_1 - \overline{Y}_2$.

There are two cases:

Equal Variances A level L confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \pm \hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2)t_{n_1+n_2-2,\frac{1+L}{2}},$$

where

$$\hat{\sigma}_p(\overline{Y}_1 - \overline{Y}_2) = \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is the estimated standard error of the point estimator $\overline{Y}_1 - \overline{Y}_2$, and

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

is the pooled variance estimator of the common population variance σ^2.

Unequal Variances A level L confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \pm \hat{\sigma}(\overline{Y}_1 - \overline{Y}_2) t_{\nu, \frac{1+L}{2}},$$

where ν is the largest integer less than or equal to

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}},$$

and

$$\hat{\sigma}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

is the estimated standard error of the point estimator $\overline{Y}_1 - \overline{Y}_2$.

TYU 28

Comparing Two Population Proportions: The Approximate Score Interval

Example 6:

In a recent survey on academic dishonesty simple random samples of 200 female and 100 male college students were taken. 26 of females and 26 of the males agreed or strongly agreed with the statement "Under some circumstances academic dishonesty is justified." Researchers would like to compare the population proportions, p_f of all female and p_m of all male college students who agree or strongly agree with this statement.

We know that a point estimate of p_f is the sample proportion $\hat{p}_f = 26/200 = 0.13$, and a point estimate of p_m is the sample proportion $\hat{p}_m = 26/100 = 0.26$. Therefore it makes sense to estimate the difference $p_f - p_m$ with the difference in estimates $\hat{p}_f - \hat{p}_m = 0.13 - 0.26 = -0.13$.

The General Case

In general, suppose there are two populations: population 1, in which a proportion p_1 have a certain characteristic, and population 2, in which a proportion p_2 have a certain (possibly different) characteristic. We will use a sample of size n_1 from population 1, and n_2 from population 2 to estimate the difference $p_1 - p_2$.

Specifically, if Y_1 items in sample 1 and Y_2 items in sample 2 have the characteristic, then $Y_1 \sim b(n_1, p_1)$, and $Y_2 \sim b(n_2, p_2)$. The sample proportion having the population 1 characteristic is $\hat{p}_1 = Y_1/n_1$, and the sample proportion having the population 2 characteristic is $\hat{p}_2 = Y_2/n_2$. A point estimator of $p_1 - p_2$ is then $\hat{p}_1 - \hat{p}_2$.

The Approximate Score Interval

As with the one sample case, recent research suggests that an interval known as the approximate score interval performs well for both large and small samples. Therefore, I will introduce it here and ask you to use it in place of the large sample (Wald) interval presented in the text. For SAS users, the macro *bici* computes the score interval (along with the exact, large-sample (Wald) and yet another interval known as the bootstrap interval (presented in Chapter 11 of the text)).

Suppose the observed values of Y_1 and Y_2 are y_1 and y_2. To compute the level L approximate score interval, first compute the adjusted estimates of p_1:

$$\tilde{p}_1 = \frac{y_1 + 0.25z^2_{(1+L)/2}}{n_1 + 0.5z^2_{(1+L)/2}},$$

and p_2:

$$\tilde{p}_2 = \frac{y_2 + 0.25z^2_{(1+L)/2}}{n_2 + 0.5z^2_{(1+L)/2}},$$

The approximate score interval for $p_1 - p_2$ is then given by the formula:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}$$

Example 6, Continued:

Recall the academic dishonesty survey in which 26 of the 200 female college students surveyed and 26 of the 100 male college students surveyed agreed or strongly agreed with the statement "Under some circumstances academic dishonesty is justified." With 95% confidence estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with this statement.

Since $z_{0.975} = 1.96$, $y_f = 26$, $n_f = 200$, $y_m = 26$, and $n_m = 100$, the adjusted estimates of p_f and p_m are

$$\tilde{p}_f = \frac{26 + 0.25 \cdot 1.96^2}{200 + 0.5 \cdot 1.96^2} = 0.1335,$$

and

$$\tilde{p}_m = \frac{26 + 0.25 \cdot 1.96^2}{100 + 0.5 \cdot 1.96^2} = 0.2645.$$

The approximate score interval for $p_f - p_m$ is then

$$0.1335 - 0.2645 \pm$$

$$1.96\sqrt{\frac{0.1335(1 - 0.1335)}{200} + \frac{0.2645(1 - 0.2645)}{100}}$$
$$= (-0.2295, -0.0325).$$

TYU 29

Tolerance Intervals

Tolerance intervals are used to give a range of values which, with a pre-specified confidence, will contain at least a pre-specified proportion of the measurements in the population.

Example 7:

Refer again to the programming time example, specifically the times obtained using the first programming language. The researchers want to obtain an interval that they are 90% confident will contain 95% of all times in the population. This is called a level 0.90 tolerance interval for a proportion 0.95 of the population values.

Mathematical Description of Tolerance Intervals

Suppose T_1 and T_2 are estimators with $T_1 \leq T_2$, and that γ is a real number between 0 and 1. Let $A(T_1, T_2, \gamma)$ denote the event $\{$*The proportion of measurements in the population between* T_1 *and* T_2 *is at least* $\gamma\}$.

Then the interval $(T_1,\ T_2)$ is a level L tolerance interval for a proportion γ of the population if

$$P(A(T_1, T_2, \gamma)) = L.$$

Normal Theory Tolerance Intervals

If we can assume the data are from a normal population, a level L tolerance interval for a proportion γ of the population is given by

$$\overline{Y} \pm KS,$$

where $\overline{Y}$ and S are the sample mean and standard deviation, and K is a mathematically derived constant depending on n, L and γ (Found in Table A.8, p. 913 in the book).

Example 7, Continued:

Refer again to the programming time example, specifically the times obtained using the first programming language. The mean of the $n = 12$ times is 17.9167, and the standard deviation is 3.6296. We checked the data and found no evidence of nonnormality. For a level 0.90 normal theory tolerance interval for a proportion 0.95 of the data, the constant K is 2.863. The interval is then

$$(17.9167 - (2.863)(3.6296),\ 17.9167 + (2.863)(3.6296))$$

$$= (7.5252,\ 28.3082).$$

TYU 30

What's the IDEA?

Note how a tolerance interval differs from both a confidence interval (a range of plausible values for a model parameter) and a prediction interval (a range of plausible values for a new observation): A tolerance interval gives a range of values which plausibly contains a specified proportion of the entire population. Tolerance intervals are confusing at first, but the following way of thinking about them may help.

Suppose we use a sample of size n to construct a level L tolerance interval for a proportion γ of a population. Call this interval 1. Now think about taking an infinite number of samples, each of size n, from the population. For each sample, we calculate a level L tolerance interval for a proportion γ of a population, using the same formula we used for interval 1. Number these intervals 2, 3, 4, ... A certain proportion of the population measurements will fall in each interval: let γ_i denote the proportion for interval i. Some of the γ_i will be greater than or equal to γ, the minimum proportion of the population measurements the interval is supposed to contain, and some will be less. The proportion of all intervals for which $\gamma_i \geq \gamma$, equals L, the confidence level.

Demo Time!

Recap:

- The models: C+E and binomial
- Types of inference: estimation, prediction, tolerance interval
- Estimation basics:
  - o Estimator or estimate?
  - o Sampling distributions
  - o Confidence intervals
- Estimation for the one population C+E model
- Prediction for the one population C+E model
- Estimation for the one population binomial model
- Determination of sample size
- Estimation for the two population C+E model
- Estimation for the two population binomial model
- Normal theory tolerance intervals