

## Introduction to Probability-Based Statistical Inference

This session is "MMD&S Ch 6 plus some." It introduces the basic ideas of using probability in the making of conclusions from data, including an introduction to

- confidence intervals
- prediction intervals (not presented in MMD&S)
- significance testing

1

The particular formulas presented here will be mathematically OK, but of limited practical use. Better ones that are both mathematically OK and practically useful will be presented next session. The ones presented here have the pedagogical value of being "simple" in light of what we have already done, and this session should be treated as simply the most straightforward introduction possible to the basic concepts.

## Introduction to Confidence Intervals

**Confidence intervals** are data-based intervals meant to bracket some unknown population or model characteristic that carry a probability-based reliability or "confidence" figure.

2

The basic idea in developing confidence interval formulas is to make use of sampling distributions for variables that involve the quantity of interest. For example, we will here see that the normal distribution of  $\bar{x}$  (guaranteed either because the population being sampled is large or because  $n$  is large) can be used to make confidence intervals for a population/universe/process/model mean,  $\mu$ .

(One notational note: it is common to now drop the convention of capitalizing random variables. So from here on, we will make no notational distinction between a random variable like the sample mean and a possible value of that random variable.)

Believing that  $\bar{x}$  may be treated as normal (either because the sample size is large or the population sampled has a normal distribution), one has the picture

3

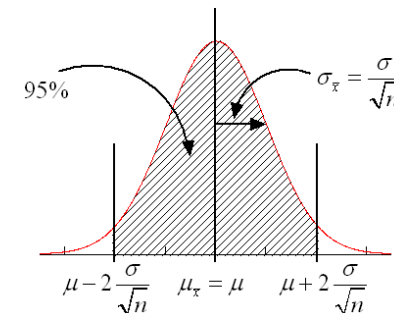


Figure 1: Probability Distribution for  $\bar{x}$

4

and about 95% of all samples will produce  $\bar{x}$ 's within  $2\sigma/\sqrt{n}$  of  $\mu$ . That means that about 95% of all samples will produce  $\bar{x}$ 's so that the interval

$$\left( \bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}} \right)$$

lands on top of  $\mu$ . This suggests that we might call

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$$

**95% confidence limits for  $\mu$ .** (Notice that these limits involve  $\sigma$ , which will rarely be known when  $\mu$  needs to be estimated.)

*Example* Suppose that I'm interested in the average rent for a 2BR apartment in Ames this coming rental year. Suppose further that historical information (perhaps from a census of rental units last year) suggests that  $\sigma = 80$ . If today a sample of  $n = 29$  apartments yields  $\bar{x} = 688.20$ , 95% confidence limits for  $\mu$  would be what?

5

Use the endpoints

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$$

which here amount to

$$688.20 \pm 2\frac{80}{\sqrt{29}}$$

that is

$$688.20 \pm 29.7$$

One can be in some sense "95% sure" that the mean rental figure is in the interval

$$(658.5, 717.9)$$

The general version of this is that confidence limits for  $\mu$  are

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

6

where  $z$  is chosen so that the area under the standard normal density curve between  $-z$  and  $z$  is the desired confidence (method reliability) figure. A small table of appropriate values of  $z$  is below (note that 2 has been replaced by a slightly more precise figure 1.96).

Confidence Level	$z$
80%	1.282
90%	1.645
95%	1.96
99%	2.576

*Demonstration* Let's illustrate estimation of the mean of the brown bag. (The fact is that the brown bag is approximately normal with  $\mu = 5$  and  $\sigma = 1.715$ .) For sake of illustration, let's make some 80% confidence intervals for  $\mu$  based on samples of size  $n = 5$ . These will have endpoints

$$\bar{x} \pm 1.282\frac{1.715}{\sqrt{5}}$$

7

i.e.

$$\bar{x} \pm .98$$

Sample	Values	$\bar{x}$	Endpoints	Successful?
1				
2				
3				
4				
5				

Notice that the "80%" figure attaches to the *methodology*, not to a specific realized interval. It is a "lifetime batting average" for the method. Each realized interval is in reality either "100% right" or "100% wrong" ... it is not "80% right" (the fact that in a real application we don't know whether it has succeeded or failed does not change this). After sample selection there is no probability left in the problem.

8

The demonstration provides a concrete way of thinking about what a confidence level means. Some applications succeed, some do not. What can be guaranteed is the lifetime batting average. Here is a set of language that is useful as "the authorized interpretation" of a confidence interval:

To say  $3 \pm .7$  are a 90% confidence limits for  $\mu$  is to say that in obtaining them I've used a method that will work in about 90% of repeat applications. Whether it has worked or not in this particular application is unknown. Regardless of this uncertainty, there is no "probability" left in the problem. (In particular, it is *not* correct to write anything like  $P(2.3 < \mu < 3.7)$ , as  $\mu$  is *not* a random variable, but rather a fixed unknown constant.)

If an application of a 90% confidence interval formula produces  $3 \pm .7$  for limits,

9

the kind of language used to describe the outcome is to say one is

95% "sure" that  $2.3 < \mu < 3.7$

or

95% "confident" that  $2.3 < \mu < 3.7$

but one *never* says that the probability is 90% that  $2.3 < \mu < 3.7$ .

*Exercise* Problem 6.11, page 377 MMD&S

MMD&S call the quantity

$$z \frac{\sigma}{\sqrt{n}}$$

the "margin of error" for estimating  $\mu$ . In some sense, a sample mean  $\bar{x}$  is "good to within" this margin of error for representing an unknown  $\mu$ . It should be clear that

10

- the larger is the confidence level, the larger is this margin of error
- the larger is the sample size, the smaller is this margin of error
- the larger is the population/universe/process/model variation ( $\sigma$ ), the larger is this margin of error

If one is somehow in advance furnished with the value of  $\sigma$ , it is possible to identify a sample size required to produce both a desired margin of error and a large confidence level by setting margin

$$\text{desired margin of error} = z \frac{\sigma}{\sqrt{n}}$$

and solving for  $n$ .

11

*Example* Ames average 2BR apartment rental example. If one, for example, were to use an historical value of  $\sigma = 80$  for planning a sample survey of this year's rental rates, and set a target margin of error for estimating  $\mu$  at, say, \$25, for 99% confidence one would need to solve

$$25 = 2.576 \frac{80}{\sqrt{n}}$$

for  $n$ . This is

$$\begin{aligned} 25\sqrt{n} &= (2.576) 80 \\ \sqrt{n} &= \frac{(2.576) 80}{25} \\ n &= \left( \frac{(2.576) 80}{25} \right)^2 \approx 68 \end{aligned}$$

The general version of the calculation in the above example produces the sample

12

size formula for estimating a mean

$$n = \left( \frac{z\sigma}{\text{desired margin of error}} \right)^2$$

## Introduction to Prediction Intervals

**Prediction intervals** are data-based intervals meant to bracket one additional observation drawn from the population/universe/process/model under consideration,  $x_{\text{new}}$ .

If one had complete information about a population of interest, the prediction problem would simply be a probability problem. (It is the fact that one only

13

has partial information, drawn from a sample, that makes this a problem of statistics rather than probability.)

*Example* Consider the brown bag (that is approximately normal with  $\mu = 5$  and  $\sigma = 1.715$ ). The limits

$$5 \pm 1.96(1.715)$$

have a 95% chance of catching  $x_{\text{new}}$ , the next value drawn from the bag. But if I don't know  $\mu$  and have to replace  $\mu$  with  $\bar{x}$ , it seems clear that I must somehow further "hedge" the prediction to account for the fact that  $\bar{x}$  only approximates  $\mu$ . The prediction interval methodology does that hedging.

To develop prediction limits, we make use of a second sampling distribution (beyond that for  $\bar{x}$ ), namely a sampling distribution for the random variable

$$x_{\text{new}} - \bar{x}$$

14

As it turns out, if the population being sampled is adequately modeled as normal (this is not a requirement that goes away on the basis a large sample size), this difference has a fairly simple sampling distribution.  $x_{\text{new}} - \bar{x}$  is itself normally distributed, with mean

$$\mu_{x_{\text{new}} - \bar{x}} = 0$$

and standard deviation

$$\sigma_{x_{\text{new}} - \bar{x}} = \sigma \sqrt{1 + \frac{1}{n}}$$

This distribution is pictured as

15

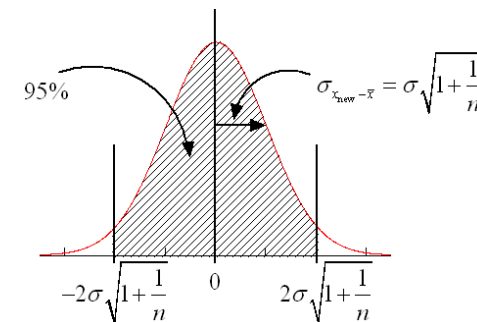


Figure 2: Sampling Distribution for  $x_{\text{new}} - \bar{x}$

16

The sampling distribution of  $x_{\text{new}} - \bar{x}$  is then such that for 95% of all experiences sampling first  $n$  values and computing  $\bar{x}$  and then selecting one more value at random ( $x_{\text{new}}$ ),

$$0 - 1.96\sigma\sqrt{1 + \frac{1}{n}} < x_{\text{new}} - \bar{x} < 0 + 1.96\sigma\sqrt{1 + \frac{1}{n}}$$

that is, 95% of all experiences will produce

$$\bar{x} - 1.96\sigma\sqrt{1 + \frac{1}{n}} < x_{\text{new}} < \bar{x} + 1.96\sigma\sqrt{1 + \frac{1}{n}}$$

This then suggests that I call

$$\bar{x} \pm 1.96\sigma\sqrt{1 + \frac{1}{n}}$$

95% prediction limits for  $x_{\text{new}}$ . In general (for prediction confidence levels

other than 95%) prediction limits for  $x_{\text{new}}$  are

$$\bar{x} \pm z\sigma\sqrt{1 + \frac{1}{n}}$$

*Example* Consider again the Ames 2BR apartment rental scenario, and now not estimation of the mean rental rate, but rather the prediction of the rental rate of one more apartment beyond the  $n = 29$  that compose a sample having  $\bar{x} = 688.20$ . With 95% confidence, prediction limits for  $x_{\text{new}}$  are

$$688.20 \pm 1.96(80)\sqrt{1 + \frac{1}{29}}$$

that is,

$$688.20 \pm 162.70$$

Notice that, of course, these limits are far wider than the 95% confidence limits for  $\mu$ . They are intended to bracket a new observation, that has its own variability  $\sigma = 80$  to consider, not just uncertainty about the value of  $\mu$ .

*Demonstration* Let's illustrate prediction of an additional individual value drawn from the brown bag based on samples of size  $n = 5$ . (Recall that that in fact the brown bag is approximately normal with  $\mu = 5$  and  $\sigma = 1.715$ .) For sake of illustration, let's make some 80% prediction intervals. These will have endpoints

$$\bar{x} \pm 1.282(1.715)\sqrt{1 + \frac{1}{5}}$$

i.e.

$$\bar{x} \pm 2.41$$

Sample	Values	$\bar{x}$	Endpoints	$x_{\text{new}}$	Successful?
1					
2					
3					
4					
5					

The "80%" guarantee attached to the prediction interval methodology illustrated above is a "lifetime batting average" guarantee that refers to the whole process of selecting  $n$  values and computing the interval, then seeing if 1 more value falls into the interval.

The demonstration clearly illustrates that in principle (unlike what is the case for a confidence interval for  $\mu$ ) one can "see" whether a particular application of the prediction interval formula works by comparing the empirical  $x_{\text{new}}$  with the prediction limits. Notice also that as sample size increases, while the length

of a confidence interval for  $\mu$  will shrink to 0, this is not true of a prediction interval for  $x_{new}$ .

*Exercise* For the situation of Problem 6.11, page 377 of MMD&S, make a 95% prediction interval for one additional study time.

## Introduction to Significance (or Hypothesis) Testing

**Significance testing** is a methodology for attaching a quantitative (probability-based) assessment of plausibility to some statement about a population/universe/process/model parameter. It operates on the basis of two "hypotheses."

21

A **null hypothesis** is a statement about a population parameter of the form

$$H_0: \text{parameter} = \#$$

that represents a "status-quo" or "pre-data" view of the situation of interest.

An **alternative hypothesis** is a statement about the population parameter of the form

$$H_a: \text{parameter} \begin{matrix} > \\ \neq \# \\ < \end{matrix}$$

that embodies the departures from  $H_0$  that are of interest (that one wishes to detect). In some contexts, the alternative hypothesis is called the "motivated hypothesis," as an investigator sets up an accepted theory as the null hypothesis, and hopes to be able to produce data making it implausible and his or her own hypothesis ( $H_a$ ) more believable.

22

*Example* Consider the filling of 20 ounce pop bottles in a beverage bottling plant. A null hypothesis about the mean fill level of bottles in the plant is

$$H_0: \mu = 20$$

A consumer advocate might be interested in the alternative hypothesis

$$H_a: \mu < 20$$

while a plant manager might employ

$$H_a: \mu \neq 20$$

*Example* Consider again the Ames 2BR apartment rental scenario. Suppose that by standards of 1) last year's mean rental and 2) this year's increase in the CPI, a mean rental rate of \$680 would be considered "fair" by a consumer

23

advocacy group. In this context, appropriate hypotheses (from the point of view of the advocacy group) might be

$$H_0: \mu = 680 \quad \text{and} \quad H_a: \mu > 680$$

Significance testing is about assessing the plausibility of  $H_0$  in view of  $H_a$ .

A **test statistic** is the data summary to be used in assessing the plausibility of  $H_0$ .

The ***p*-value** in a significance testing problem is the probability that the sampling distribution possessed by the test statistic if  $H_0$  is true assigns to values "more extreme" than the one actually observed.

*Example* Consider again the Ames 2BR apartment rental scenario, and significance testing from the point of view of the consumer group. An obvious test

24

statistic for an hypothesis about  $\mu$  is the sample mean,  $\bar{x}$ . The situation can be pictured as below, where the  $p$ -value is the probability that a sample mean of  $n = 29$  observations from a population with mean  $\mu = 680$  (the value from the null hypothesis) exceeds 688.20 (the observed value of the test statistic).

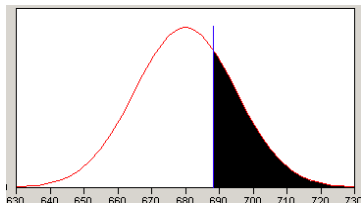


Figure 3: The Distribution of  $\bar{x}$  Under the Null Hypothesis Has Mean 680 and Standard Deviation  $80/\sqrt{29} = 14.86$

To get a  $p$ -value in this problem, we need to compute an area under the normal curve, and for that a  $z$ -score is needed. Here, this is

$$\begin{aligned} z &= \frac{\bar{x} - \#}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{688.20 - 680}{\frac{80}{\sqrt{29}}} \\ &= .55 \end{aligned}$$

Notice that we subtract the mean and divide by the standard deviation *corresponding to the distribution of  $\bar{x}$  under the null hypothesis*. A table look-up gives .7088 as the corresponding left-tail area, so

$$p\text{-value} = 1 - .7088 = .2919$$

This says that outcomes (values of  $\bar{x}$ ) more extreme (in this case, larger than) the one I've seen would occur 29% of the time if  $H_0$  were true. That fails to

make  $H_0$  implausible, and the consumer group doesn't have iron-clad evidence of a mean rental rate that is above the "fair" value of \$680.

"Small"  $p$ -values correspond to "strong evidence against  $H_0$ " (in the direction of the specified alternative hypothesis). They make  $H_0$  "implausible."

"Large"  $p$ -values correspond to "weak evidence against  $H_0$ " (in the direction of the specified alternative hypothesis). They do not make  $H_0$  "implausible."

People often use language like "Assess the strength of the evidence that XXXX." in stating significance testing problems. When they do, "XXXX" belongs in the alternative hypothesis. Similarly, when they ask "Is there statistically detectable evidence that XXXX?" the alternative hypothesis should embody XXXX.

In the Ames rent example, we used

Hypotheses	Test Statistic	$p$ -value
$H_0: \mu = \#$ $H_a: \mu > \#$	$\bar{x}$	a right-tail normal area

but in order to actually compute a  $p$ -value, we had to convert  $\bar{x}$  to a  $z$ -score. In fact, we might as well say that for testing  $H_0: \mu = \#$ , the test statistic will be the  $z$ -score corresponding to  $\bar{x}$ , namely

$$z = \frac{\bar{x} - \#}{\frac{\sigma}{\sqrt{n}}}$$

Then for the 3 different possible alternative hypotheses, we get  $p$ -values as

Alternative Hypothesis	$p$ -value
$H_a: \mu < \#$	a left-tail standard normal area
$H_a: \mu > \#$	a right-tail standard normal area
$H_a: \mu \neq \#$	the sum of right- and left-tail standard normal areas

This can be pictured (for \* marking an observed value of the test statistic) as

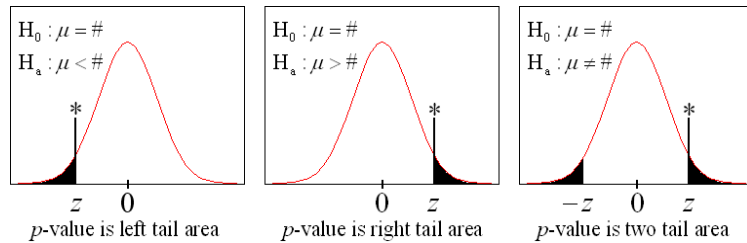


Figure 4:  $p$ -values for Three Alternative Hypotheses

### Comments/Philosophy/Caveats Regarding Significance Testing

By many standards, significance testing is much less informative than is confidence interval estimation. Nevertheless, it is common and must be taught so that you can communicate and understand what people are talking about when it is used. Here are some cautions about the real application and interpretation of significance testing.

- Reasonably common terminology is that
  - a  $p$ -value less than .05 is a "statistically significant" result
  - a  $p$ -value less than .01 is a "highly statistically significant" result

But you should be careful not to hear in this language something that is *not true*. That is,

statistical significance  $\neq$  practical importance

To have statistical significance is to have enough data to definitively see that  $H_0$  is incorrect. That is *not* the same as  $H_0$  being far enough from correct that it makes a practical difference! Statistical significance is sample size dependent. Given a large enough sample, *any* departure from  $H_0: \mu = \#$  can be detected and statistical significance achieved.

- A  $p$ -value is *not* a "probability that  $H_0$  is true (or that it is false)." It is a measure of implausibility of the null hypothesis, nothing more, nothing less.

- It can be argued that significance tests attempt to answer the "wrong questions." They ask things like "Is  $\mu = 17$ ?" And the truth is that almost certainly,  $\mu$  is not *exactly* 17. Maybe 17.000498567, but not 17. The issue is not whether the mean is exactly 17 but rather "What is  $\mu$ ?" And it is *confidence interval estimation* that attempts to answer this ("right") question, not significance testing.

- Not only does confidence interval estimation attempt to answer the "right" question, but it provides some significance testing information "for free." If, for example, a 99% confidence interval for  $\mu$  is

(5.2, 7.3)

(and thus doesn't contain the number 8.0) the  $p$ -value for testing  $H_0: \mu = 8.0$  versus  $H_a: \mu \neq 8.0$  is less than  $1 - .99 = .01$ .