April 28, 2005 Lecture Notes

Measurement Error, Probability, Models for Means

**Measurement error**: No matter how carefully made, any measurement, if repeated, could come out differently. Why?

**Basic measurement error equation**:

\[
\text{(each individual measurement)} = \text{(exact "true" value)} + \text{(bias)} + \text{(a random error)}
\]

Systematic error = 0 in this case study

Random errors fluctuate around 0 – mean of random errors is 0 + SD = 0.2

**Case Study 9**:

\[
\text{(each individual measurement)} = \text{(exact "true" value)} + \text{(bias)} + \text{(a random error)}
\]

\[
\begin{align*}
3.7 &= \text{potassium (K) reading 1} = 3.8 + 0 + (-0.1) \\
4.0 &= \text{potassium (K) reading 2} = 3.8 + 0 + (+0.2)
\end{align*}
\]

Unobservable mean of random errors is 0

**K** = symbol for potassium

**Population**

- Conceptual: all possible K readings for you

**K Concentration**

\[
\begin{bmatrix}
\vdots \\
\vdots \\
\end{bmatrix}
\]

**Sample**

- The observed K readings

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n \\
\end{bmatrix}
\]

Mean \( \bar{Y} \) = ? (ex: 3.6)

**Imaginary data set**

\[
\begin{bmatrix}
3.6 \\
4.0 \\
\vdots \\
\end{bmatrix}
\]

Need to know:

- Long run mean: 3.8
- Long run SD: 0.1

**Population histogram**

- SD = 0.2
- (follows normal curve)

**Histogram of \( Y \)** (one draw)

- SD = 0.2
- \( 3.5 - 3.8 = -1.5 \) (87% according to chart)

100\% - 87\% = 13\% chance of misdiagnosis

- 2 of 100
\[ n = 4 \text{ (we want 4 readings)} \]

\[ P(\text{misdiagnosis } w/n = 4) = P(\bar{y} < 3.5 \text{ with } \bar{y} \text{ based on } n = 4 \text{ readings}) \]

**Long Run Mean** = Expected value of \( \bar{y} = E(\bar{y}) = EV = E_{100}(\bar{y}) = \mu = 3.8 \)

**Long Run SD** = Standard Error of \( \bar{y} = SE(\bar{y}) = SE = SE_{100}(\bar{y}) \)

(noise of uncertainty of \( \bar{y} \) as estimate of \( \mu \))

\[ N \] - nothing to do w/SE
\[ \mu \] - nothing to do w/SE
\[ \sigma \] - \( \sigma \uparrow \) \( SE(\bar{y}) \uparrow \)
\[ n \] - \( n \uparrow \) \( SE(\bar{y}) \downarrow \)

**most important formula in stats:**
\[
SE(\bar{y}) = \frac{\sigma}{\sqrt{n}}
\]

(uncertainty about \( \mu \) goes down as \( n \uparrow \) but only at a rate of \( \frac{1}{\sqrt{n}} \); to cut SE in half, I need to quadruple \( x4 \) sample size \( n \))

thus, \[
SE_{100}(\bar{y}) = \sigma \uparrow \sqrt{n} \quad SE_{100}(\bar{y}) = \frac{\sigma}{\sqrt{n}} \uparrow \sqrt{n}
\]

\[
SE(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{\sqrt{4}} = 0.1
\]

The **Long Run SD** is 0.1

**Long Run Histogram of \( \bar{y} \) w/n = 4:**

\[ SE = 0.1 \]

\[ SD = 0.2 \]

3.5 - 3.8 = -3

100% - 97% = 0.15% 

\[ 0.1 \]

2 chance of

according to chart on p.23, misdiagnosis 99.7% in middle

6.5% chance of misdiagnosis (from previous page)
<table>
<thead>
<tr>
<th>Cost</th>
<th>n (# of tests)</th>
<th>P (incorrect diagnosis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$25</td>
<td>1</td>
<td>6.5%</td>
</tr>
<tr>
<td>$100</td>
<td>4</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

Here, the consequences of misdiagnosis are relatively mild (eating bananas when you don’t need to).

**Decision theory: utility:** What you attach to correct diagnosis.

(you decide if it is worth the extra $75 to have a better chance of correct diagnosis.)

(End of Case Study 9)

Why does it help to take an average of (n>1) readings instead of just 1?

**No bias:**

observation 1 = truth + random error 1
observation 2 = truth + random error 2

\[ \text{observation } n = \text{truth} + \text{average of } n \text{ random errors} \]

ex: \((-0.1) + (0.2) + \ldots + (-0.3) + (0.1)\) = cancelation of \(\text{+} \Theta \text{ errors} \)

\[ 1/n \text{ measures likely size of mean of } n \text{ errors} \]

**W/ bias:**

observation 1 = truth + bias + random error 1

\[ \text{observation } n = \text{truth} + \text{bias} + \text{average of } n \text{ random errors} \]

\[ \text{gets to 0 as } n \to \infty \]

however, bias will increase, making data worse

so, w/ bias, (average of n observations) = truth + bias (which only gets worse)
Beginning of Case Study 10:

**Population**
- All 22,984 waybills
- Amount owed:
  - Mean $\mu = ?$
  - SD $\sigma = ?$

**Sample**
- Observed waybills
- Amount owed:
  - Mean $\bar{y} = 28.09$
  - SD $s = 31.40$

Using SRS method: $n = 2,072$