Analysis of Experimental/Sample Data

4/30/12

Note Title

Population – large – mean μ and variance σ^2 (estimate?)

– a $\underbrace{\text{small subset of the population}}_{\text{sample}}$ is used to estimate the pop. parameters

with or without replacement

– Each person's response is a random variable

Sample size is n

a random sample is a collection of values for the n RVs

$X_1, \ldots, X_n$, each representing a person in the sample

Let the sample values be $x_1, x_2, \ldots, x_n$

So for this sample, $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \quad \text{is a RV.}$$

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n} E(X_1 + \cdots + X_n)$$

$$= \frac{1}{n}\left[E(X_1) + \cdots + E(X_n)\right], \quad E(X_1) = \cdots = E(X_n) = \mu$$

$$= \frac{1}{n} \cdot \{n\mu\} = \mu$$

$\boxed{V(aX) = a^2 V(X)}$

$$V(\bar{X}) = V\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\left[V(X_1 + \cdots + X_n)\right] = \frac{1}{n^2}\left[V(X_1) + \cdots + V(X_n)\right]$$

[X_i's are indep.

$$= \frac{1}{n^2} \cdot n\sigma^2 = \boxed{\frac{\sigma^2}{n}}$$

Summarizing the central tendency of sample data

To summarize experimental and sample data by a single number, a central tendency measure is often used.

Suppose a sample of size n is to be summarized. The commonly used central tendencies are

Mean – average of data

Median – the middle # of data: $\lceil \frac{n}{2} \rceil$th # if n is odd; av. of $\lceil \frac{n}{2} \rceil$th and $\lceil \frac{n}{2} \rceil + 1$ # if n is even

Mode – the most freq. value in the data

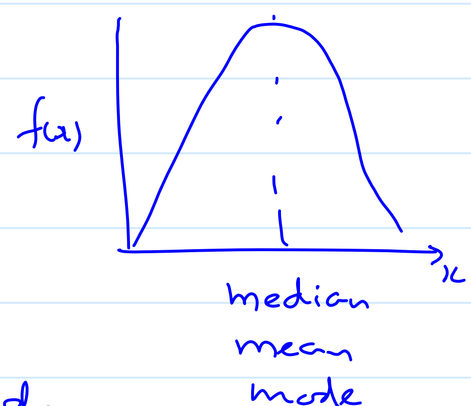
Example: Consider the sample: 3,8,2,11,6

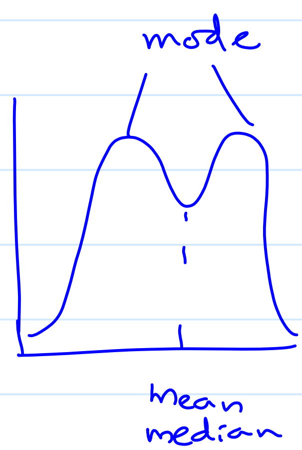
Here n = 5.
The mean is (3+8+2+11+6)/5 = 6.0

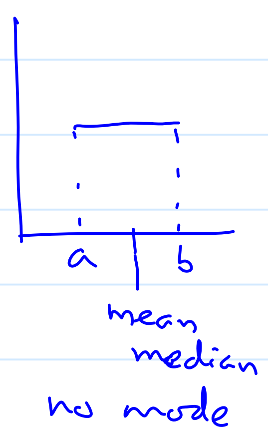
Median: First sort the sample data: 2, 3, 6, 8 ,11
The $\lceil n/2 \rceil$th value of the sorted sequence is the median. In this example, it is the 3rd value which is 6.

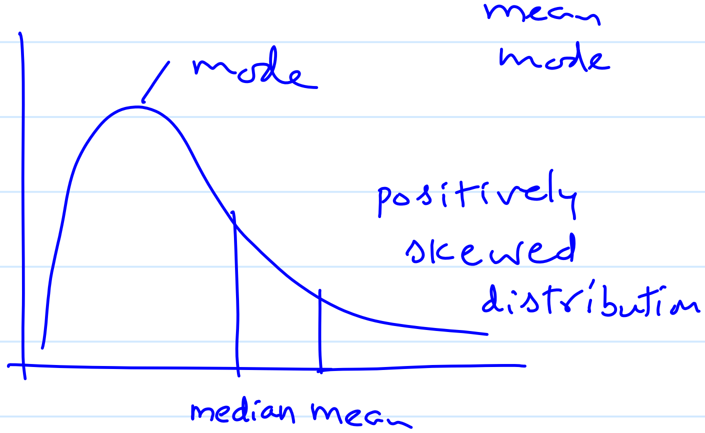
Mode: there is no mode since there is no value that is occuring more frequently than the others.

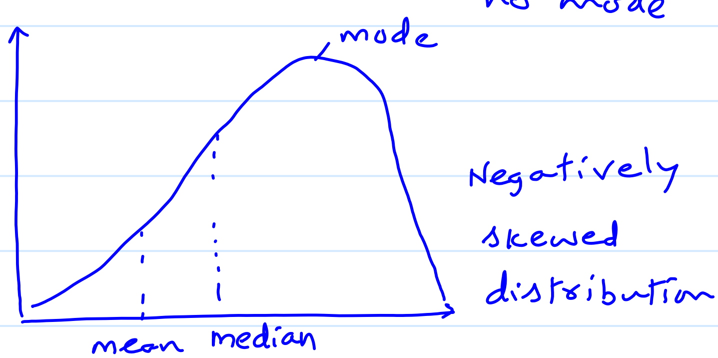
Examples











Summarizing the variability of sample data

Range – min & max values of sample data

Variance or standard deviation of sample data

Quantiles and percentiles

SIQR (semi-interquartile range)

Mean absolute deviation

Let $x_1, x_2, \dots, x_n$ be the sorted values of the sample

mean $\bar{x} = \dfrac{x_1 + x_2 + \dots + x_n}{n}$

Variance $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

s = standard deviation.

For a uniform RV U that takes values $u_1, \dots, u_n$ with equal prob. $\frac{1}{n}$,

$V(U) = \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^2$

median is $\lceil \frac{n}{2} \rceil$th value of the sorted sample, if n odd

or av. of $\frac{n}{2}$ and $\frac{n}{2}+1$ values if n is even

Quartiles and Percentiles

Take the sorted sample. Let they be $x_1, x_2, \ldots x_n$.

α-quantile of this sample is $\lceil n\alpha \rceil^{th}$ value of the sorted sequence

It is often denoted as x_α.

kth value in the sorted has the quantile $\left(\frac{k-0.5}{n}\right)$.

$x_{0.5}$ is the median = 0.5 quantile value

Percentile: α quantile is 100α percentile.

Quartiles are $x_{0.25}, x_{0.5}, x_{0.75}$

Quintiles are $x_{0.2}, x_{0.4}, x_{0.6}, x_{0.8}$

deciles are $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$

SIQR : $\dfrac{x_{0.75} - x_{0.25}}{2}$

Mean absolute deviation (MAD) is calculated as $\dfrac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$

Example: 3, 8, 2, 11, 6 $\xrightarrow{\text{Sort}}$ 3, 8, 2, 11, 6

$n=5$

$$\text{mean} = \frac{3+8+2+11+6}{5} = 6 = \bar{x}$$

Range: max − min values

$11-2$

$$\text{Variance } s^2 = \frac{1}{n-1} \sum_{i=1}^{5} (x_i - \bar{x})^2$$

$$= \frac{1}{4}\left((2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2\right)$$

$$= \frac{1}{4}(54) = 13.5$$

$$\text{std. dev. } s = \sqrt{13.5} = 3.67$$

i	x_i	Quantiles	Percentiles
1	2	$\frac{1-0.5}{5} = 0.1$	10
2	3	$\frac{2-0.5}{5} = 0.3$	30
3	6	$\frac{3-0.5}{5} = 0.5$	50
4	8	$\frac{4-0.5}{5} = 0.7$	70
5	11	$\frac{5-0.5}{5} = 0.9$	90

What is the 65 percentile value?

$\Rightarrow$ What is the 0.65 quantile value?

$x_{0.65}$ is $\lceil 5 * 0.65 \rceil = 4^{th}$ value

$x_{0.65} = 8$

Confidence Intervals

Sample data $x_1, x_2, \ldots, x_n$

$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ is the sample mean

$s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample variance

For sufficiently large n, say n>= 30, and under some conditions,

$$\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} , \quad \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right) \quad — (1)$$

gives the interval in which the true mean (μ) of the population resides with prob. (1-α).

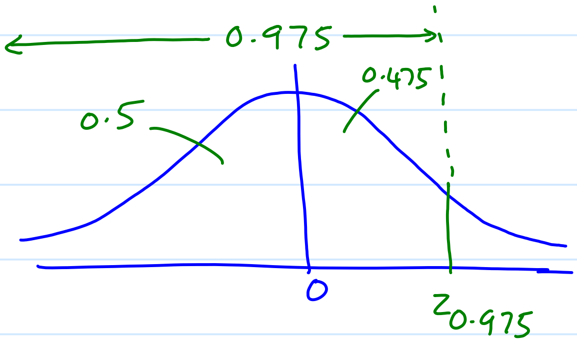
The interval in (1) is called the confidence interval on the estimation of μ by the sample mean (xbar) and 100*(1-α) is the confidence level.

Often, 95% confidence level is used, for which α = 0.05.

$z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the standard normal RV.

For 95% confidence level, $\alpha = 0.05$ $[\because 100(1-\alpha) = 95$

$z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$



E = error estimate $= z_{1-\frac{\alpha}{2}} \cdot \dfrac{s}{\sqrt{n}}$ C.I. is $(\bar{x} - E, \bar{x} + E)$.

Example

For the sample of 2, 3, 6, 8, 11, calculate the 95% confidence interval for the sample mean.

$\bar{x} = 6 \qquad s = 3.67 \qquad n = 5$

95% confidence level $\Rightarrow$ $100(1-\alpha) = 95 \Rightarrow \alpha = 0.05$

Error estimate, $E = Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = Z_{0.975} \cdot \frac{s}{\sqrt{n}}$

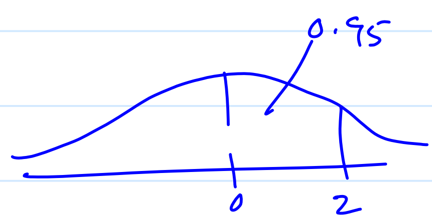
$$= 1.96 * \frac{3.67}{\sqrt{5}} = 3.2$$

$\therefore$ C.I. is $\underbrace{(6-3.2,\ 6+3.2)}_{(2.8,\ 9.32)}$ at 95% confidence level.

what is the 90% confidence interval?

$100(1-\alpha) = 90 \Rightarrow 1-\alpha = 0.9 \Rightarrow \alpha = 0.1$

$Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.64$



$E = Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 1.64 * \frac{3.67}{\sqrt{5}} = 2.7$

C.I. $(6-2.7,\ 6+2.7) = (3.3,\ 8.7)$

What is 99% confidence interval?

$100(1-\alpha) = 99 \Rightarrow \alpha = 0.01$ $\qquad Z_{1-\alpha/2} = Z_{0.995} = \text{Table}^{-1}(0.495) = 2.57$

E = 2.57 * 3.67/sqrt(5) = 4.2

C.I. is (6 -+ 4.2) or (1.8, 10.2) Too wide to be meaningful.

C.I. can be narrowed by reducing the confidence level and/or by increasing the number of samples.

Q. from Spring 2011 Final:

To assess the consumer sentiment, a random collection of 100 people are asked to rate their outlook on the future prospects on a scale of 1 to 100. Higher numbers mean more optimistic. From this sample, it is determined that the average consumer sentiment score is 75 with a standard deviation of 10. Calculate the 95% confidence interval for the average consumer sentiment score.

$\bar{x} = 75, \quad s = 10, \quad \alpha = 0.05, \quad n = 100$

$C.I. = \bar{x} \mp E, \qquad E = z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 1.96 \cdot \frac{10}{\sqrt{100}} = 1.96$

$\therefore$ C.I. is $(75 \mp 1.96) = (73.04, 76.96)$