Statistical Inference

Statistical inference is using data analysis to infer properties of a probability distribution.

It assumes that a sampled data set can be used to make predictions about a larger population from which the sample is drawn.

what is sampling?

A selection of elements from a finite or infinite population

why sample?

1) Speed: quicker
2) Cost: cheaper
3) Accuracy: tendency to miscount for large numbers
4) Necessity: can't test for quality on every item if they will be destroyed e.g. matches

how should samples be drawn?

Always randomly, a sample is random when each element in the population has an equal chance of being selected. There is a lack of bias or predictability.

Note: a random sample is not necessarily a cross-section of the population

Notation/Terminology

	Population	**Sample**
collection method	census	survey

Notation/Terminology

	Population	**Sample**
collection method	census	survey
$P(X = x)$	probability $p(x)$	relative frequency f_r

Notation/Terminology

	Population	**Sample**
collection method	census	survey
$\mathrm{P}(X = x)$	probability $p(x)$	relative frequency f_r
$\mathrm{E}(X)$	μ	$\overline{x}$

Notation/Terminology

	Population	**Sample**
collection method	census	survey
$P(X = x)$	probability $p(x)$	relative frequency f_r
$E(X)$	μ	$\overline{x}$
$Var(X)$	σ^2	s^2

Notation/Terminology

	Population	**Sample**
collection method	census	survey
$P(X=x)$	probability $p(x)$	relative frequency f_r
$E(X)$	μ	$\bar{x}$
$Var(X)$	σ^2	s^2
Note: as $n \to \infty$; $\bar{x} \to \mu$ and $s^2 \to \sigma^2$		

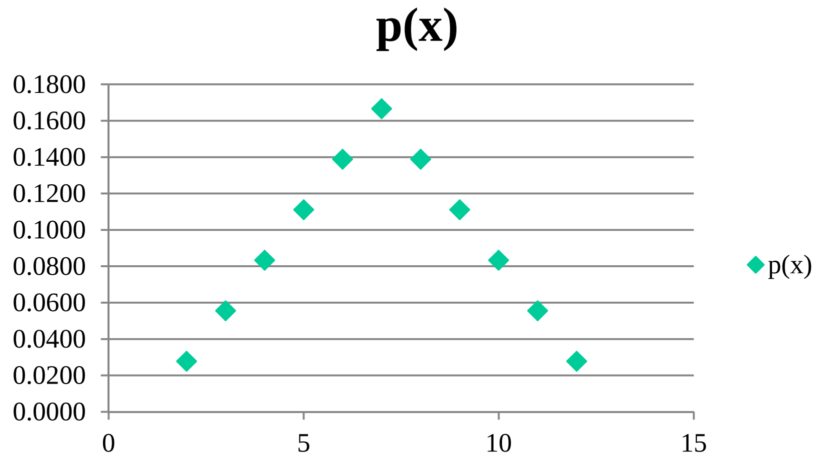
e.g. Two dice are rolled. Use the theoretical probability distribution to calculate $E(X)$ and $Var(X)$ for the population

x	2	3	4	5	6	7	8	9	10	11	12	Sum
p(x)	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
xp(x)	0.0556	0.1667	0.3333	0.5556	0.8333	1.1667	1.1111	1.0000	0.8333	0.6111	0.3333	7.0000
x2p(x)	0.1111	0.5000	1.3333	2.7778	5.0000	8.1667	8.8889	9.0000	8.3333	6.7222	4.0000	54.8333

$$E(X) = 7$$

$$\begin{aligned} Var(X) &= E(X^2) - \mu^2 \\ &= 54.8\dot{3} - 49 \\ &= 5.8\dot{3} \end{aligned}$$

$$\sigma = 2.4152$$



(ii) Roll a pair of dice 50, 100, 500, 1000 times and compare these samples with the population.

50

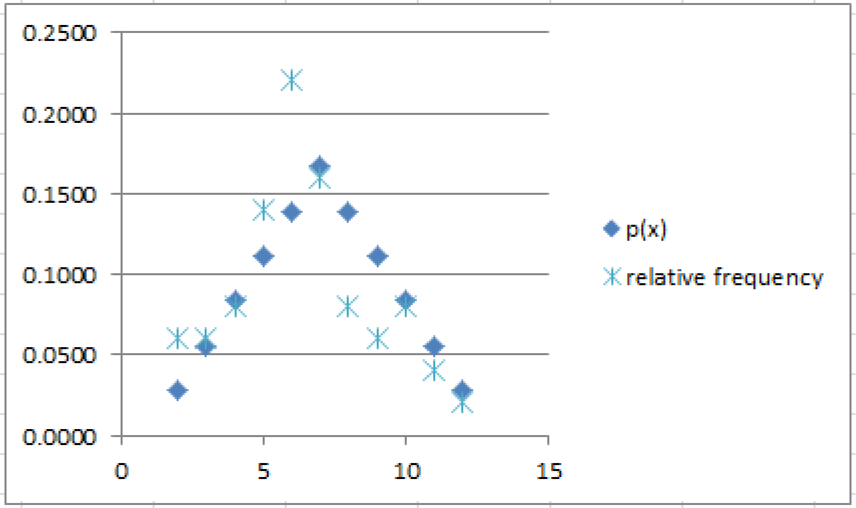
x	2	3	4	5	6	7	8	9	10	11	12	Sum
p(x)	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
xp(x)	0.0556	0.1667	0.3333	0.5556	0.8333	1.1667	1.1111	1.0000	0.8333	0.6111	0.3333	7.0000
x2p(x)	0.1111	0.5000	1.3333	2.7778	5.0000	8.1667	8.8889	9.0000	8.3333	6.7222	4.0000	54.8333
Frequency	3	3	4	7	11	8	4	3	4	2	1	50
relative frequency	0.0600	0.0600	0.0800	0.1400	0.2200	0.1600	0.0800	0.0600	0.0800	0.0400	0.0200	1.0000
x times fr	0.1200	0.1800	0.3200	0.7000	1.3200	1.1200	0.6400	0.5400	0.8000	0.4400	0.2400	6.4200
x2p(x)	0.2400	0.5400	1.2800	3.5000	7.9200	7.8400	5.1200	4.8600	8.0000	4.8400	2.8800	47.0200

$$\mathrm{E}(X) = \underline{6.42} \qquad |\mu - \overline{x}| = 0.58$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \overline{x}^2$$

$$= 47.02 - 6.42^2$$

$$= \underline{5.8036} \qquad |\sigma - s| = 0.0061$$

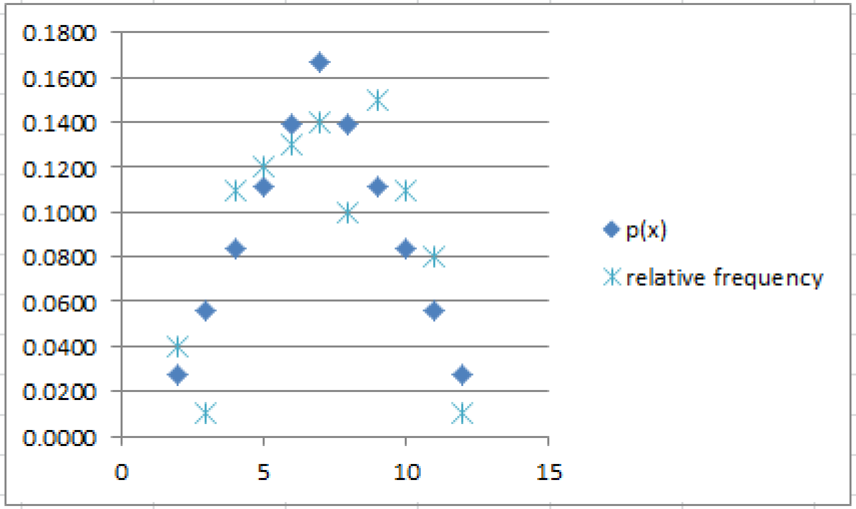


100

x	2	3	4	5	6	7	8	9	10	11	12	Sum
$p(x)$	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
$xp(x)$	0.0556	0.1667	0.3333	0.5556	0.8333	1.1667	1.1111	1.0000	0.8333	0.6111	0.3333	7.0000
$x2p(x)$	0.1111	0.5000	1.3333	2.7778	5.0000	8.1667	8.8889	9.0000	8.3333	6.7222	4.0000	54.8333
Frequency	4	1	11	12	13	14	10	15	11	8	1	100
relative frequency	0.0400	0.0100	0.1100	0.1200	0.1300	0.1400	0.1000	0.1500	0.1100	0.0800	0.0100	1.0000
x times fr	0.0800	0.0300	0.4400	0.6000	0.7800	0.9800	0.8000	1.3500	1.1000	0.8800	0.1200	7.1600
$x2p(x)$	0.1600	0.0900	1.7600	3.0000	4.6800	6.8600	6.4000	12.1500	11.0000	9.6800	1.4400	57.2200

$$\mathrm{E}(X) = \underline{7.16} \qquad |\mu - \overline{x}| = 0.16$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \overline{x}^2 \\ &= 57.22 - 7.16^2 \\ &= \underline{5.9544} \qquad |\sigma - s| = 0.0250\end{aligned}$$

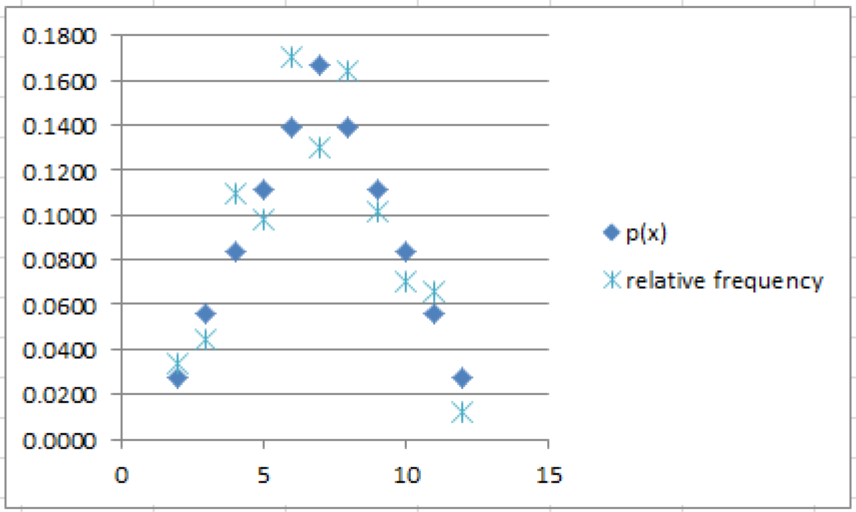


500

x	2	3	4	5	6	7	8	9	10	11	12	Sum
p(x)	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
xp(x)	0.0556	0.1667	0.3333	0.5556	0.8333	1.1667	1.1111	1.0000	0.8333	0.6111	0.3333	7.0000
x2p(x)	0.1111	0.5000	1.3333	2.7778	5.0000	8.1667	8.8889	9.0000	8.3333	6.7222	4.0000	54.8333
Frequency	17	22	55	49	85	65	82	51	35	33	6	500
relative frequency	0.0340	0.0440	0.1100	0.0980	0.1700	0.1300	0.1640	0.1020	0.0700	0.0660	0.0120	1.0000
x times fr	0.0680	0.1320	0.4400	0.4900	1.0200	0.9100	1.3120	0.9180	0.7000	0.7260	0.1440	6.8600
x2p(x)	0.1360	0.3960	1.7600	2.4500	6.1200	6.3700	10.4960	8.2620	7.0000	7.9860	1.7280	52.7040

$$\mathrm{E}(X) = \underline{6.86} \qquad |\mu - \bar{x}| = 0.14$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \bar{x}^2 \\ &= 52.704 - 6.86^2 \\ &= \underline{5.6444} \qquad |\sigma - s| = 0.0394\end{aligned}$$

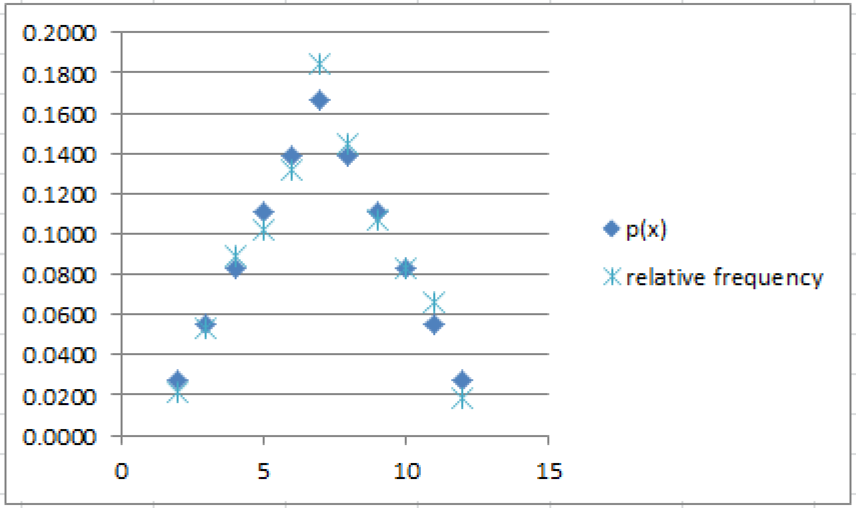


1000

x	2	3	4	5	6	7	8	9	10	11	12	Sum
p(x)	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278	1
xp(x)	0.0556	0.1667	0.3333	0.5556	0.8333	1.1667	1.1111	1.0000	0.8333	0.6111	0.3333	7.0000
x2p(x)	0.1111	0.5000	1.3333	2.7778	5.0000	8.1667	8.8889	9.0000	8.3333	6.7222	4.0000	54.8333
Frequency	21	53	89	102	132	184	145	107	83	66	18	1000
relative frequency	0.0210	0.0530	0.0890	0.1020	0.1320	0.1840	0.1450	0.1070	0.0830	0.0660	0.0180	1.0000
x times fr	0.0420	0.1590	0.3560	0.5100	0.7920	1.2880	1.1600	0.9630	0.8300	0.7260	0.2160	7.0420
x2p(x)	0.0840	0.4770	1.4240	2.5500	4.7520	9.0160	9.2800	8.6670	8.3000	7.9860	2.5920	55.1280

$$\mathrm{E}(X) = \underline{7.042} \qquad |\mu - \bar{x}| = 0.042$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \bar{x}^2 \\ &= 55.128 - 7.042^2 \\ &= \underline{5.5382} \qquad |\sigma - s| = 0.0619\end{aligned}$$



Exercise 13D; 1, 2, 10, 11