1 Data Summarization

1.1 Distributions

The data distrib describes prob. of random value taking a particular value

E.g: IQ follows Normal (Gaussian) distrib.

Random no. generator generates random nos. that follow uniform distribution in [0,1]

A roll of a biased dice has a categorical distrib. over the values $\{1, 2, 3, 4, 5, 6\}$

1.2 Measures of central tendency

Sample statistic is a measurement on a sample from a distrib. calculated by applying a function to the sample

Measures of central tendency are sample stats that attempt to capture where middle of distrib is.

Three common ways of measureing central tendency

- Mean
- Median
- Mode

1.3 Mean, Median, Mode

- Arithmetic Mean:
  - Sum of observations divided by no. of observations

$$\overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

- Median:
  - Middle value that separates higher and lower half of dataset
  - Sort data and take mid value
  - Even number of observations: take arithmetic mean of middle two values
  - Robust Statistic (less sensitive to outliers)
- Mode:
  - Most freq. value in dataset
  - Suitable measure of central tendency for nominal variables
  - Easy to compute for discrete values
  - Not straightforward for continuous distrib.

	Dimension	Type	Calculate
Mean	N	Quantitative	np.mean
Median	1	Ordinal	np.median
Mode	N	Nominal	From histogram

1.4 Sample Mean vs. Population Mean

Population mean is trye mnean for entire pop.: μ

Sample mean is calculated mean from sample of pop: $\overline{x}$

E.g. pop. mean human height would require measuring height of every person on earth. Sample mean 100 randomly chosen people

As sample size increases, sample mean approaches pop. mean

$$\lim_{n\to\infty} \overline{x}_n = \mu.$$

1.5 Measures of Statistical Dispersion

Measure how stretched/squeezed a distrib is

Most common methods:

- Variance σ^2
- Std. Dev σ
- Interquartile Range (IQR)

Can be defined for sample or pop.

1.6 Variance, SD

- Variance
  - Expected squared dev from mean

$$\delta^2 = E[(X - EX)^2].$$
$$= E[(X - \mu)^2].$$
$$\hat{\sigma}^2 \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Usually don't know pop mean μ, so we just use sample mean:

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

For small samples, this is biased, use bias corrected estimator:

$$\hat{\sigma}_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Above known as Bessel's correction

Std. dev is sqrt of (possibly bias corrected) variance

$$\delta = \sqrt{\sigma^2}.$$

1.7 Range and interquartile range

Range of sample is diff. between min and max values

Range is not robust stat

IQR is range in which 50% of data lies

$$IQR = Q_3 - Q_1.$$

1.7.1 Interquartile Range

IQM can be used on ordinal vars and often makes more sense than the std. dev

1.8 Sufficient Statistics

Stat. is sufficient wrt stat model and assoc. unknown paoram when no other stat that can be calculated from the same sample provides any additional info as tp the val of the param

For Normal distribution, sufficient stats are mean and std. dev

$$= \{\mu, \sigma\}.$$

Once sufficient stats. known you know everything there is to know about distrib of var.

1.9 Skewness and Kurtosis

Most data not normally distrib.

Skewness: measure of assymmetry of prob. distrib. of real values random var about its mean

Kurtosis: Measure of how heavy tails are of prob. distrib of real-valued random var

- Pearsons moment coefficient of skewness:
  - Third standardized moment:

$$_1 = E[(\frac{X - \mu}{\sigma})^3].$$

- Pearson's moment coefficient of kurtosis:
  - Fourth standardized moment:

$$_2 = E[\frac{X - \mu}{\sigma)^4}].$$

- Replace mu and sigma with sample stat.s to calculate for sample

1.10 Conditioning on Categorical Variable

Conpute stats based on subsets of sample

Subsets selected by grouping on categorical vars

These are stats on the conditions distrib.

1.11 Statistics of association between attributes

Covariance: how two vars vary together

$$cov(X, Y) = E[(X - \mu_x)(Y - x_y)].$$

Pearson product-moment correlation coefficient (Pearson's rho):

- Measure of linear relationship between quantitative vars
- Normalized version of the covariance

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

- Ranges from -1 to 1
- >0: + correlation
- <0: - correlation
- +1 indicates perfect linear relationshiop between X and Y
- -1 is perfect negative correlation

Not suitable for ordinal variables

1.12 Spearman Correlation

Spearmans rank correlation coefficient (Spearman's rho):

- Pearson correlation between rank values of those two vars

Pearson's correlation assesses linear relationships, Spearmans correlation assess monotonic relationships (whether linear or not)

2 Data Visualization

2.1 Why Visualize Data?

- Explore
  - Understand data and relationships between attribs.
  - Generate new hypothese
  - Detect errors and anomalies
- Confirm
  - Hypotheses already generated
  - Check assumptions
  - Verify conclusions
- Communicate
  - Analysis has been completed
  - Conclusions verified
  - Task: Present and Communicate results
  - Inform, persuade, educate, entertain

2.2 Statistical Data Visualization

Provide a set of standard visual tools for:

- Undestanding the shape and distrib. of data
- Understanding relationships between attribs.
- Understanding composition and part-whole relationships
- Comparing quantities