Descriptive Statistics

Damien Benveniste

October 15, 2017

1 Sample Moments

1.1 The Mean

The first raw moment:

Definition

$$\mu = E[X] = \sum_{x \in X} xP(x) \tag{1}$$

Estimation

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

1.2 The Variance

The second centered moment:

Definition

$$\sigma^2 = E[(X-\mu)^2] = E[X^2] - E[X]^2 \tag{3}$$

Estimation

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 \tag{4}$$

1.3 The Skewness

The third central normalized moments:

Definition

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{5}$$

Estimation

$$G = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2\right]^{3/2}} \tag{6}$$

1.4 The Kurtosis

The fourth central normalized moments:

Definition

$$\kappa = \frac{E[(X-\mu)^4]}{\sigma^4} \tag{7}$$

Estimation

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2\right]^{2}} \tag{8}$$

2 Bias/unbiased estimators

The bias of an estimator is the difference the estimator's expected value and the true value of the parameter being estimated. We assume $\{X_1, X_2, \ldots, X_n\}$ to be independent and identically distributed (i.i.d.) random variables with expected value μ and variance σ^2. Let's consider the estimator of the mean $M = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\begin{aligned} E[M] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned} \tag{9}$$

Therefore this estimator is unbiased. Let's consider now the estimator of the variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$:

$$
\begin{aligned}
E[S^2] &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - M)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu + \mu - M)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2 + 2(X_i - \mu)(\mu - M) + (\mu - M)^2] \\
&= \sigma^2 - E[(\mu - M)^2] \\
&= \sigma^2 - Var[M]
\end{aligned} \tag{10}
$$

We have

$$
\begin{aligned}
Var[M] &= Var[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \\
&= \frac{\sigma^2}{n}
\end{aligned} \tag{11}
$$

Therefore

$$
\begin{aligned}
E[S^2] &= \sigma^2 - Var[M] \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned} \tag{12}
$$

Then S^2 is biased but $\frac{n}{n-1}S^2$ is not.

3 Quantiles

q-Quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes. x is a k^{th} q-quantile for a variable X if

$$
Pr[X < x] \leq \frac{k}{q} \tag{13}
$$

with $0 < k < q$. Important quantiles:

- Median: $q = 2$
- Deciles: $q = 10$
- Quartiles: $q = 4$
- Percentiles: $q = 100$
- ...

4 Measures of dependency

4.1 Pearson correlation

Linear dependency

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{14}$$

4.2 Spearman's rank correlation

Have the variables the same order?

$$r_{X,Y} = \rho_{rg_X, rg_Y} = \frac{E[(rg_X - \mu_{rg_X})(rg_Y - \mu_{rg_Y})]}{\sigma_{rg_X} \sigma_{rg_Y}} \tag{15}$$

Example: $X = \{1, 6, 2, 9, 0\}$ and $Y = \{0, 60, 1, 500, -8\}$. We get $rg_X = \{2, 4, 3, 5, 1\}$ and $rg_Y = \{2, 4, 3, 5, 1\}$, therefore $r_{X,Y} = 1$

4.3 Information Theory

4.3.1 Entropy

Expected amount of information:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \tag{16}$$

4.3.2 Mutual Information

How dependent are variables?

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right) \tag{17}$$

If $I(X;Y) \geq 0$ and if $I(X;Y) = 0$ there is no mutual information of the variables are independent

4.3.3 Kullback-Leibler divergence

How similar are distributions?

$$D(P\|Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{18}$$