Properties of Expected Values (Expectations), III

3. $\mathrm{E}\,(b_1 + X_1) = b_1 + \mathrm{E}\,(X_1)$.

4. $\mathrm{E}\,\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}\,(X_i)$, n is finite.

5. $\mathrm{E}\,\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i\, \mathrm{E}\,(X_i)$.

$$E\left(\sum_{i=1}^{n} \frac{1}{n} X_i\right) = \sum_{i=1}^{n} \frac{1}{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} E(X_i).$$

Covariance

Assume that $\mathrm{E}(|X_1|)$, $\mathrm{E}(|X_2|)$ and $\mathrm{E}(|X_1 X_2|)$ are all finite. Then the *covariance* between X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1, X_2) = E(X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2)).$$

An alternative equivalent definition is (by linerity of expectations).

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1 X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2).$$

bilinearity of covariances: let a_1 and a_2 be const's

$$\mathrm{Cov}(a_1 X_1, a_2 X_2) = a_1 \cdot a_2 \cdot \mathrm{Cov}(X_1, X_2)$$

Symmetry: notice $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.

Main property: bilinearity of covariances

Proposition: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be rvs with well-defined covariances $Cov(X_i, Y_j)$ for every i and j. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ be any constants. Then

$$Cov(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j). \quad (\star)$$

Let C be a matrix such that $C_{ij} = Cov(X_i, Y_j)$

Let $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$; $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$. $a^T = (a_1, \ldots, a_n)$.

Then $(\star) = a^T \cdot C \cdot b$

$= b^T \cdot C^T \cdot a$

Variance is a special case of covariance

<u>*Variance of a rv.*</u> If $X_1 = X_2$, then

$$Cov(X_1, X_2) = Cov(X_1, X_1) = \mathrm{E}\left(X_1^2\right) - \left(\mathrm{E}\left(X_1\right)\right)^2 = Var(X).$$

$$= E\left(\left[X_1 - E(X_1)\right]^2\right) \geq 0$$

Q: Is there a difference between $\mathrm{E}\left(X^2\right)$ and $\left(\mathrm{E}\left(X\right)\right)^2$?

A: yes. Since $Var(X_1) \geq 0$, $E(X^2) \geq (E(X_1))^2$

<u>*Variance of a linear combination of rvs*</u>

$$Var\left(\sum_{j=1}^{n} a_j X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \cdot a_j \cdot Cov(X_i, X_j)$$

using matrix algebra, $= a^T \cdot C \cdot a$

Random variables X_1 and X_2 are said to be <u>*uncorrelated*</u> if $Cov(X_1, X_2) = 0$.

Correlation

Correlation coefficient (loosely, *correlation*) between rvs X_1 and X_2 that have a finite second moment is defined as

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}.$$

Let $\rho = Corr(X_1, X_2)$. If $\rho = 0$, the rvs are *uncorrelated*.

Q: Why is $|\rho| \leq 1$?

Independence and correlation

Let X_1 and X_2 be rv's with the joint cdf $F_{1,2}$, joint pdf/pmf $f_{1,2}$, marginal cdf's F_1 and F_2 and marginal pdf's/pmf f_1, f_2.
Recall that *X_1 and X_2 are independent* if and only if
$F_{1,2}(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$ if and only if
$f_{1,2}(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$ for every x_1, x_2.

Independence of X_1 and X_2 does not guarantee existence of moments.
However if $\mathrm{E}\left(X_1^2\right) < \infty$ and $\mathrm{E}\left(X_2^2\right) < \infty$ and X_1 and X_2 are independent, we have

$$\star \qquad \mathrm{E}\left(X_1 \cdot X_2\right) = \mathrm{E}\left(X_1\right) \mathrm{E}\left(X_2\right) \Rightarrow \mathrm{Cov}\left(X_1, X_2\right) = 0,$$

i.e., X_1 and X_2 are uncorrelated.

$$E(X_1 X_2) = \int\int x_1 \cdot x_2 \cdot \underbrace{f_{12}(x_1, x_2)}_{f_1(x_1) \cdot f_2(x_2) \text{ by indep.}} dx_1\, dx_2 = \left(\int x_1 f_1(x_1)\, dx_1\right) \cdot \left(\int x_2 \cdot f_2(x_2)\, dx_2\right) = E(X_1) \cdot E(X_2)$$

$\mathrm{Cov}\left(X_1, X_2\right) = 0$ *does not imply that X_1 and X_2 are independent.*

Parametric family of distributions

<u>Notation:</u> $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is the parametric family of distributions indexed by θ.

<u>Parameterization</u> is the correspondence between θ and F_θ.

<u>Common setup</u>: $\underbrace{X_1, X_2, \ldots, X_n}_{\text{iid sample from } F_\theta} \overset{\text{iid}}{\sim} F_\theta$.

WMS: $X_1, X_2, \ldots, X_n$ is a "random sample" from F_θ.

Warning: in general, "random" does not mean "independent".

Goals of probability and statistics

Goal of probability: determine $\Pr(T(X_1, \ldots, X_n) \in A)$, where T is some function.

To motivate **goals of statistics**, consider a game:

1. Mother Nature chooses $\theta \in \Theta$ and generates $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$.
2. Goal of statistics/statistician: use info in the sample $X_1, \ldots, X_n$ to make inference/learn/guess the value of θ that the nature has chosen.

What is meant by statistical inference?

Inference = estimation + hypothesis testing

Estimation:

1. Point estimation: use of data $X_1, X_2, \ldots, X_n$ to "guess" θ using a rv $T(X_1, \ldots, X_n)$ that is close to θ in some prob. sense.
2. Set/interval estimation: find a random set/interval $S(X_1, \ldots, X_n)$ such that $\Pr(\theta \in S(X_1, \ldots, X_n))$ is high.

Hypothesis testing: use the sample to determine if a hypothesis $\theta = \theta_0$ is likely to be true. "Do the data support the hypothesis that $\theta = \theta_0$?"

Other goals of statistics (besides inference): modeling, prediction.

Check out the video about Ritz Casino scam for an amazing illustration of how modeling and prediction are useful in real life
`https://www.youtube.com/watch?v=Gna0M4W-hDE`

Identifiability of a parameterization

Recall the **goal of statistics** and the game:

1. Mother Nature chooses $\theta \in \Theta$ and generates $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$.
2. Goal of statistics/statistician: use info in the sample $X_1, \ldots, X_n$ to make inference/learn/guess the value of θ that the nature has chosen, or the distribution F_θ.

A parameterization $\theta \mapsto F_\theta$ is called *identifiable* if $\theta \neq \theta'$ implies $F_\theta \neq F_{\theta'}$.

E. Non-identifiable parameterization

$F_\theta = \text{Normal}(\mu_1 + \mu_2, \sigma^2)$; $\theta = (\mu_1, \mu_2, \sigma^2)$.

μ_1: weight of owner; μ_2: weight of a pet

$\theta' = (\mu_1 + c, \mu_2 - c, \sigma^2)$ corresponds to the same distribution

In linear regression, this is known as "multicollinearity".

Estimators and estimates

Def. An *estimator* of $g(\theta)$ is a statistic $T(X_1, X_2, \ldots, X_n)$ that is used to "guess" the value of $g(\theta)$.

Def. An *estimate* of $g(\theta)$ is the value of the estimator $T(X_1, \ldots, X_n)$ when $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$.

Notice: $T(X_1, \ldots, X_n)$ is a rv (an estimator), while $T(x_1, \ldots, x_n)$ is a fixed number (an estimate).

Our interest is in finding good estimators.

Two principal methods are the method of moments and the maximum likelihood estimation (discussed later).

To compare the "goodness" estimators, it is necessary to define several criteria, below. It is assumed here that θ is a scalar; this can be generalized to the case when θ is a vector.

Bias

Def. *Bias* of an estimator $\widehat{\theta}$ of θ is

$$Bias(\theta|\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Notice that, typically, the bias is a function of θ (and, possibly, of other parameters).

If $Bias(\theta|\widehat{\theta}) = 0$ for every value of θ, the estimator $\widehat{\theta}$ is called *unbiased*.

Examples:

MSE and its bias-variance decomposition

Def. *Mean squared error* of an estimator $\widehat{\theta}$ is

$$MSE(\theta|\widehat{\theta}) = E\{(\theta - \widehat{\theta})^2\}.$$

Bias-variance decomposition of the MSE:

$$MSE(\theta|\widehat{\theta}) = \{Bias(\theta|\widehat{\theta})\}^2 + Var(\widehat{\theta}).$$

Consequence: bias-variance tradeoff.