Variance

Lemma 1: Suppose $X_1, \ldots, X_n$ be a sample (i.i.d., SRS, or other). Then

$$\sum_i (X_i - \overline{X})^2 \;=\; \ldots \;=\; \sum_i X_i^2 - n\overline{X}^2.$$

Proof: Expand the expression. □

Let $\mathbf{X} = \langle X_1, \ldots, X_n \rangle$, $\overline{\mathbf{X}} = \overline{X}\mathbf{1} = \overline{X}\,\langle 1, \ldots, 1 \rangle = \langle \overline{X}, \ldots, \overline{X} \rangle$, and $\mathbf{V} = \mathbf{X} - \overline{\mathbf{X}}$. Note that

$$|\mathbf{X}|^2 = \sum_i X_i^2, \qquad |\overline{\mathbf{X}}|^2 = n\overline{X}^2, \qquad \text{and} \qquad |\mathbf{V}|^2 = \sum_i (X_i - \overline{X})^2.$$

Lemma 2: Suppose $X_1, \ldots, X_n$ is an i.i.d. random sample from a population with mean μ, variance σ^2. Then the sample variance, defined as

$$S^2 := \frac{\sum_i (X_i - \overline{X})^2}{n-1}, \quad = \frac{|\vec{V}|^2}{n-1}$$

is an unbiased estimator for the population variance σ^2. That is,

$$\mathrm{E}\left(\frac{\sum_i (X_i - \overline{X})^2}{n-1} \right) = \sigma^2.$$

Proof: First, note in general that, since $\mathrm{Var}(W) = \mathrm{E}\,(W^2) - [\mathrm{E}\,(W)]^2$, we have that

$$\mathrm{E}\,(X_i^2) = \mu^2 + \sigma^2, \qquad \text{for each } i \text{, and} \qquad \mathrm{E}\,(\overline{X}^2) = \mu^2 + \frac{\sigma^2}{n}.$$

Combining this with the result of the previous lemma, we show

$$\mathrm{E}\left(\sum_i (X_i - \overline{X})^2 \right) \;=\; \mathrm{E}\left(\sum_i X_i^2 - n\overline{X}^2 \right) \;=\; \sum_i \mathrm{E}\,(X_i^2) - n\,\mathrm{E}\,(\overline{X}^2) \;=\; \cdots \;=\; (n-1)\sigma^2,$$

which leads to the claim. □

Notes:

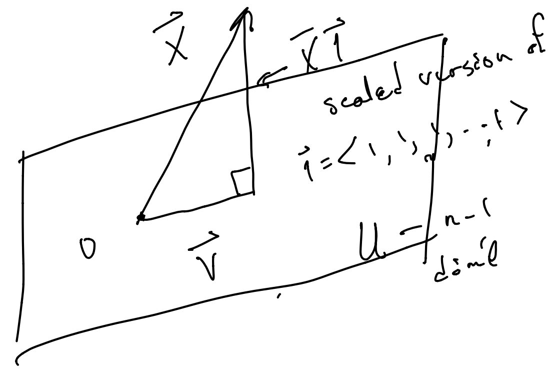
- For the seemingly "more natural" estimator

$$\hat{\sigma}^2 = \frac{\sum_i X_i - \overline{X})^2}{n}, \qquad \text{we have} \qquad \mathrm{E}\,(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2.$$

  That is, it is *biased*.

$\vec{X} = \langle x_1, x_2, \ldots, x_n \rangle$, $\overline{X} = \frac{1}{n}\sum X_i$

$\vec{\bar{X}} \cdot \vec{1}$



- Viewing the results of the lemmas above in terms of vector lengths, we have

$$(n-1)S^2 \;=\; |\mathbf{V}|^2 \;=\; |\mathbf{X}|^2 - |\overline{\mathbf{X}}|^2,$$

or

$$|\mathbf{V}|^2 + |\overline{\mathbf{X}}|^2 \;=\; |\mathbf{X}|^2 \;=\; |\mathbf{V} + \overline{\mathbf{X}}|^2.$$

Since the Pythagorean Theorem is an "if and only if" result, it follows that $\mathbf{V} \perp \overline{\mathbf{X}}$.

- We may choose an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ for $\mathbb{R}^n$ with $\mathbf{u}_1$ parallel to $\mathbf{1}$. Then

$$\mathbf{X} \;=\; \overline{\mathbf{X}} + \mathbf{V}$$

$\mathcal{U} = \text{span}\{\vec{u}_2, \ldots, \vec{u}_n\}$

where

  - $\overline{\mathbf{X}} = \text{proj}(\mathbf{X} \to \mathbf{u}_1)$
  - If we write $\mathcal{U} = \text{span}(\mathbf{u}_2, \ldots, \mathbf{u}_n)$, then

$$\mathbf{V} = \text{proj}(\mathbf{X} \to \mathcal{U}) = \sum_{i=2}^{n} \text{proj}(\mathbf{X} \to \mathbf{u}_i) = \sum_{i=2}^{n} (\mathbf{X} \cdot \mathbf{u}_i)\mathbf{u}_i,$$

the second equation a result of orthogonality (see Friday's notes). And, after repeated application of the Pythagorean theorem,

$$(n-1)S^2 = \quad |\mathbf{V}|^2 = \sum_{i=2}^{n} (\mathbf{X} \cdot \mathbf{u}_i)^2,$$

  - If $\langle X_1, \ldots, X_n \rangle \overset{\text{i.i.d.}}{\sim} \mathsf{Norm}(\mu, \sigma)$, then by defining $W_i := \mathbf{X} \cdot \mathbf{u}_i$ for $i = 2, \ldots, n$, we get the $W_i \overset{\text{i.i.d.}}{\sim} \mathsf{Norm}(0, \sigma)$, and

$$(n-1)S^2 = |\mathbf{V}|^2 \;=\; \sum_{i=2}^{n} (\mathbf{X} \cdot \mathbf{u}_i)^2 \;=\; W_2^2 + W_3^2 + \cdots + W_n^2.$$

In light of all this, we begin using a new distribution.

Definition 1: Let $\mathbf{Z} = \langle Z_1, \ldots, Z_n \rangle \overset{\text{i.i.d.}}{\sim} \mathsf{Norm}(0, 1)$. The sum

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

has the **chi-squared distribution** with n **degrees of freedom**, abbreviated as $\mathsf{Chisq}(n)$.

Using the cdf method, for $X = Z_1^2 + \cdots + Z_n^2$ as above, one can show that $X \sim \mathsf{Chisq}(n)$, then $X \sim \mathsf{Gamma}(\alpha = n/2, \lambda = 1/2)$

Case $n = 1$: $X = Z_1^2$ w/ $Z \sim \text{Norm}(0,1)$

$$F_X(x) = P_r(X \le x) = P_r(Z_1^2 \le x) = P_r(-\sqrt{x} \le Z_1 \le \sqrt{x})$$

$$= \Phi(\sqrt{x}) - \Phi(-\sqrt{x})$$

So $$f_X(x) = \frac{d}{dx}\left[\Phi(\sqrt{x}) - \Phi(-\sqrt{x})\right] = \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{x})^2/2} \cdot \frac{1}{2x^{1/2}} + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{x})^2/2} \frac{1}{2\sqrt{x}}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\cancel{2}\sqrt{x}} e^{-x/2} \cancel{(1+1)} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{x}} e^{-x/2}$$

= pdf for another dist., namely $\text{Gamma}(\alpha = 1/2, \lambda = 1/2)$

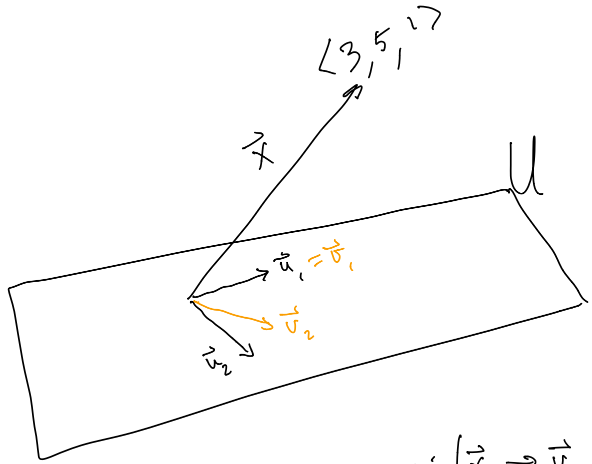
Generally: $\underbrace{Z_1^2 + \cdots + Z_n^2}_{\text{independent}} \sim \underline{\text{Gamma}}(n/2, 1/2)$

So, $\mathsf{Chisq}(n)$ is a special case of a Gamma distribution. We have introduced it because of this result:

Lemma 3 (Lemma 4.6.6, p. 267): Let $\mathbf{X} = \langle X_1, \ldots, X_n \rangle \overset{\text{i.i.d.}}{\sim} \mathsf{Norm}(\mu, \sigma)$. Then

(i) $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \sim \mathsf{Chisq}(n-1)$, and

(ii) $\overline{X}$ and S^2 are independent random variables.



$$\text{proj}(\vec{x} \to \vec{u}_1) + \text{proj}(\vec{x} \to \vec{u}_2)$$
$$= \langle 4,4,0\rangle + \langle -1,1,0\rangle$$
$$= \langle 3,5,0\rangle$$
$$\stackrel{?}{=} \text{proj}(\vec{x} \to U)$$

$$\text{proj}(\vec{x} \to \vec{v}_1) + \text{proj}(\vec{x} \to \vec{v}_2)$$
$$= \langle 4,4,0\rangle + \langle 3,0,0\rangle = \langle 7,4,0\rangle$$
$$\stackrel{?}{=} \text{proj}(\vec{x} \to U)$$

Conclusion(?): When you have an orthogonal basis for a subspace U, then

$$\text{proj}(\vec{x} \to U) = \text{sum of projections} \left(\vec{u} \to \begin{matrix}\text{basis} \\ \text{vectors}\end{matrix}\right).$$