Probability Theory

Textbook: *Introduction to Probability* by Blitzstein and Hwang

Previous Lecture

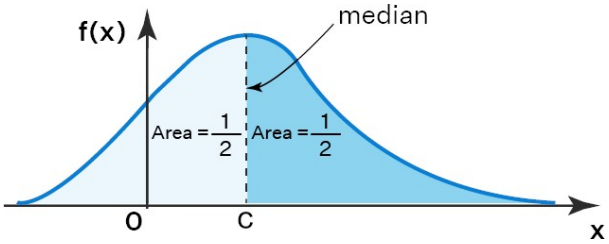
- Normal Distr $\mathcal{N}(\mu, \sigma^2)$: PF/CDF/Mean/Var
- Normal symmetry properties, standardization, empirical rule
- Exponential Distr: PF/CDF/Mean/Var
- Memoryless Property

§6.1 - Summaries of a Distribution

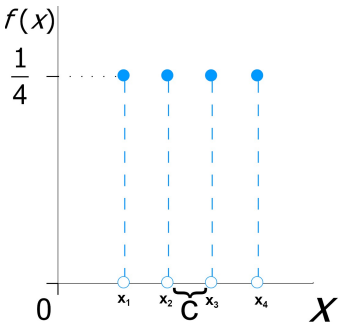
We know of some single number summaries of distr (mean, median, variance).
Where do they come from? Are there more?

Def (**Median**): We say that c is a median of X if $P(X \leq c) \geq \frac{1}{2}$ and $P(X \geq c) \geq \frac{1}{2}$.

The simplest way this can happen is if the CDF of X hits $\frac{1}{2}$ exactly at c....



but we know that some CDFs have jumps....

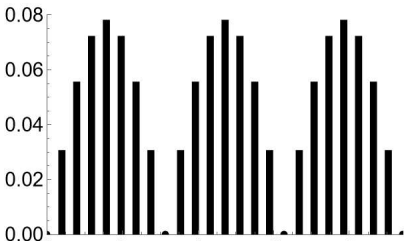


Discrete PF, CDF has jumps

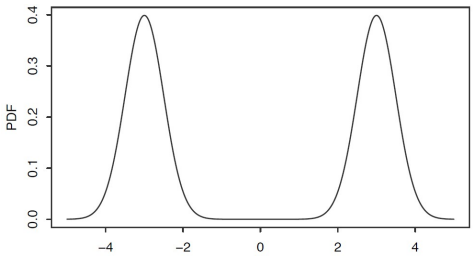
Medians between (and including) x_2 and x_3

Def (**Mode**): For discrete X, c is a mode of X if it maximizes the PF: $P(X = c) \geq P(X = x)$ for all x.

For cont X, c is a mode of X if it maximizes the PF: $f(c) \geq f(x)$ for all x.



Discrete PF: three modes



Cont PF: two modes, $X = -3, 3$

Ex (Discrete Median/Mode): Let discrete X have the PF: $f_x(x) = \begin{cases} \frac{12}{25x} \text{ for } x = 1,2,3,4, \\ 0 \quad \text{otherwise.} \end{cases}$

a. Find the least median of X.

$P(X \le 1) = F(1) = \frac{12}{25(1)} \approx 0.48$

$P(X \le 2) = F(2) = \frac{12}{25} + \frac{12}{25(2)} \approx 0.72 \ge \frac{1}{2}.$

$P(X \ge 2) = 1 - F(1) \approx 0.52 \ge \frac{1}{2}.$ So $c = 2$ is the least (and only) median of X.

b. Find the mode(s) of X.

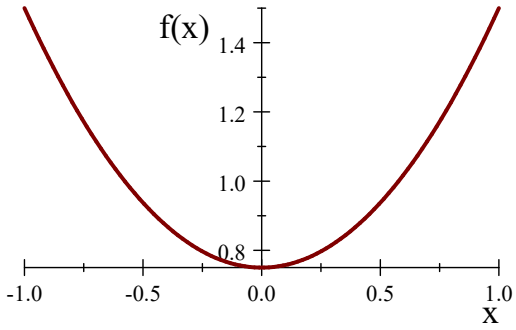
Observe that f_x decreases as x increases, so the mode of X is $x = 1$.

Ex (Cont. Median/Mode): Let cont X have the PF: $f_x(x) = \begin{cases} \frac{3}{4} + \frac{3}{4}x^2, & -1 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$

a. Find $E(X)$.

$E(X) = \int_{-1}^{1} x f(x) dx = \frac{3}{4} \int_{-1}^{1} (x + x^3) dx$ (odd funct with symmetric bounds, or...)

$= \frac{3}{4} \left[\frac{1}{2}x^2 + \frac{1}{4}x^4 \right]_{-1}^{1} = \frac{3}{4} \left(\frac{1}{2} + \frac{1}{4} - \left(\frac{1}{2} + \frac{1}{4} \right) \right) = 0.$



$\frac{3}{4} + \frac{3}{4}x^2$

b. Find the 3rd moment of X.

$$E(X^3) = \int_{-1}^{1} x^3 f(x)dx$$

$$= \tfrac{3}{4}\int_{-1}^{1} (x^3 + x^5)dx \qquad \text{(odd funct with symmetric bounds, or...)}$$

$$= \tfrac{3}{4}\left[\tfrac{1}{4}x^4 + \tfrac{1}{6}x^6\right]_{-1}^{1} = \tfrac{3}{4}\left(\tfrac{1}{4} + \tfrac{1}{6} - \left(\tfrac{1}{4} + \tfrac{1}{6}\right)\right) = 0.$$

c. How would you find the mode(s)?

Take the derivative of f_x and set it equal to zero to solve for critical pnts.

Then check if they are max or mins (either w/second derivative, or first derivative test).

Then check the endpoints of the domain.

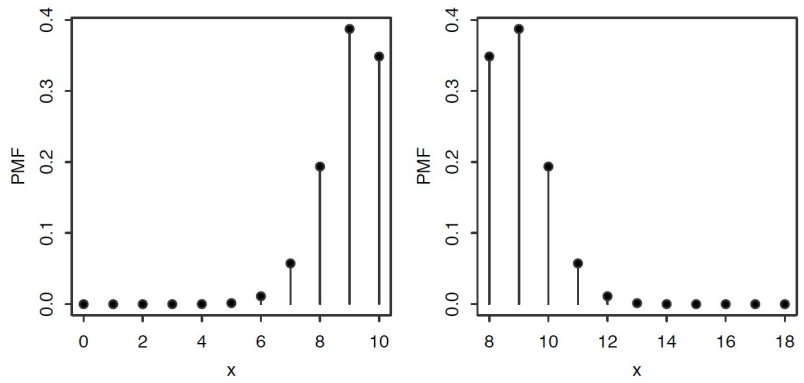
Suppose that we are trying to guess what a not-yet-observed X will be, by making a prediction c. The mean and the median both seem like natural guesses for c, but which is better? That depends on how "better" is defined. Two natural ways to judge how good c is are the mean squared error $E(X-c)^2$ and the mean absolute error $E|X-c|$. The following result says what the best guesses are in both cases.

Thm: Let X have mean μ, and m be a median.

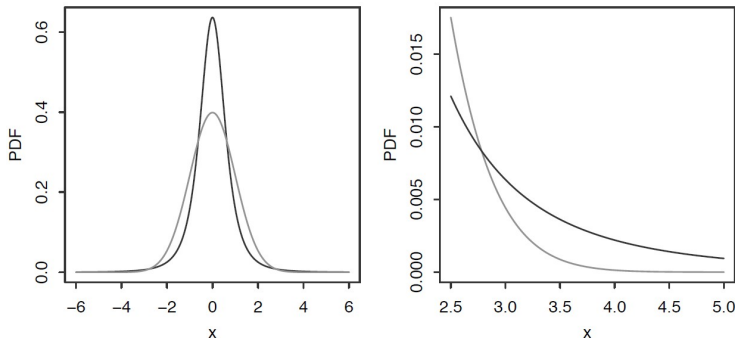
The value of c that minimizes the mean squared error $E(X-c)^2$ is $c = \mu$.

A value of c that minimizes the mean absolute error $E|X-c|$ is $c = m$.

[Proofs in Book]



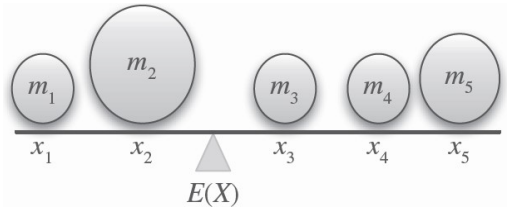
Left: $Bin(10, 0.9)$ is left-skewed. Right: $Bin(10, 0.1)$, shifted to the right by 8, is right-skewed but has the same mean, median, mode, and variance as $Bin(10, 0.9)$.



§6.2 - Interpreting Moments

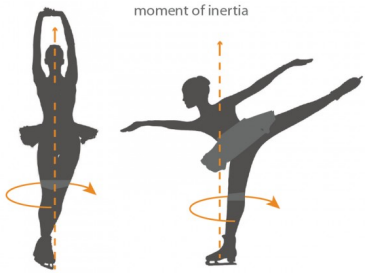
Def (Kinds of Moments): Let X have mean (**location**) μ and variance (**spread**) σ^2. For any positive integer n, the nth **moment** of X is $E(X^n)$, the nth **central moment** is $E((X-\mu)^n)$, and the nth **standardized moment** is $E\left(\left(\frac{X-\mu}{\sigma}\right)^n\right)$.

("if it exists" is implicit for all these)



In physics: $E(X) = \sum_{j=1}^{n} m_j x_j$ is called the **center of mass** of a system, and

$Var(X) = \sum_{j=1}^{n} m_j (x_j - E(X))^2$ is called the **moment of inertia** about the center of mass.



Ex (Continued from Above): Let discrete X have PF: $f_x(x) = \begin{cases} \frac{12}{25x} & \text{for } x = 1,2,3,4, \\ 0 & \text{otherwise.} \end{cases}$

c. Find the first moment of X.

$$E(X) = \sum_{x=1}^{4} x \frac{12}{25x} = 4\left(\frac{12}{25}\right) = \frac{48}{25} \approx 1.92$$

d. Find the 2nd central moment of X.

$$E\left((X-E(X))^2\right) = E\left(\left(X-\frac{48}{25}\right)^2\right) = E\left(X^2 - \frac{96}{25}X + \left(\frac{48}{25}\right)^2\right)$$

$$= E(X^2) - \frac{96}{25}E(X) + E\left(\left(\frac{48}{25}\right)^2\right) = E(X^2) - \frac{96}{25}\frac{48}{25} + \left(\frac{48}{25}\right)^2 = E(X^2) - \frac{2304}{625}.$$

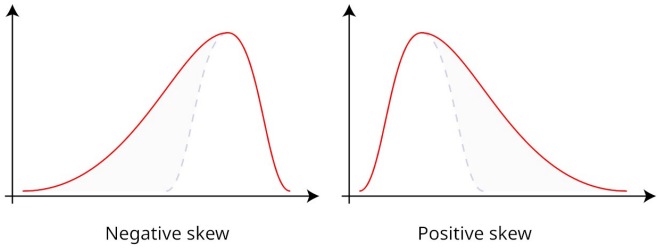
$$E(X^2) = \sum_{x=1}^{4} x^2 \frac{12}{25x} = \frac{12}{25}\sum_{x=1}^{4} x = \frac{12}{25}(1+2+3+4) = \frac{24}{5}.$$

So therefore, $E\left((X-E(X))^2\right) = \frac{24}{5} - \frac{2304}{625} = \frac{696}{625} \approx 1.11.$

Skewness

When we go up another level we find **skewness**, a single-number summary of **asymmetry**

Def (**Skewness**): The skewness of X w/mean μ and variance σ^2 is the third standardized moment of X: $Skew(X) = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right)$.



Def (**Symmetry of a Rv**): We say X has a **symmetric distr about** μ if $X-\mu$ has the same distr as $\mu - X$. We say X is **symmetric** or X has a **symmetric distr**; these have the same meaning.

Proposition (**Odd Central Moments of a Symmetric Distr**). Let X be symmetric about its mean μ. Then for any odd number m, the mth central moment $E(X-\mu)^m$ is 0 if it exists.

Proof. Since $X-\mu$ has the same distr as $\mu - X$, they have the same mth moment (if it exists):
$E(X-\mu)^m = E(\mu - X)^m$.

Let $Y := (X-\mu)^m$. Then $Y = (\mu - X)^m = (-(X-\mu))^m = (-1)^m Y = -Y$.

So the above equation just says $E(Y) = -E(Y)$. Thus, $E(Y) = 0$. ■

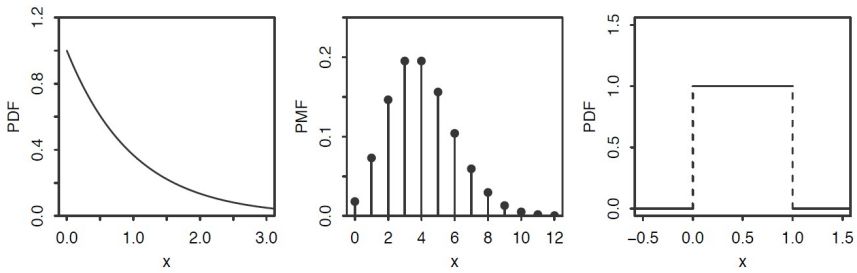
Since any odd the central moment of a symmetric distr is zero, we could have used any of them as the definition of skewness, so why did we use the 3rd central moment?

The first central moment is zero for a symmetric distr. And the higher odd central moments are more difficult to calculate, and turn out to be less accurate.

Kurtosis

When we go up another level we find **kurtosis**, a single-number summary of **the thickness of a distr's tail**.

Def (**Kurtosis**): The kurtosis of X w/mean and variance σ^2 is a shifted version of the fourth standardized moment of X: $Kurt(X) = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3$.



Skewness and kurtosis of some named distributions.

Left: $Expo(1)$ PDF, skewness = 2, kurtosis = 6.

Middle: $Pois(4)$ PMF, skewness = 0.5, kurtosis = 0.25.

Right: $Unif(0,1)$ PDF, skewness = 0, kurtosis = −1.2.

The "−3" seems (and is) somewhat arbitrary. It's not universally defined this way, but has the benefit of making the kurtosis of the Normal distr equal zero.

---

§6.3 - Sample Moments

How do we use collected (sampled) data to estimate unknown parameters of a distr?

Imagine some random process (plinko?) where you **don't know the distr**.

You can run the process (drop the plinko chips) and collect results from the iid rvs representing that process.

If the data are iid $X_1, \ldots, X_n$ where the mean $\mu = E(X_j)$ is unknown, the most obvious way to estimate the mean is to average the X_j, taking the arithmetic mean.

Def (Sample Moments). Let $X_1, \ldots, X_n$ be iid. The kth sample moment is $M_k := \frac{1}{n}\sum_{j=1}^{n} X_j^k$.

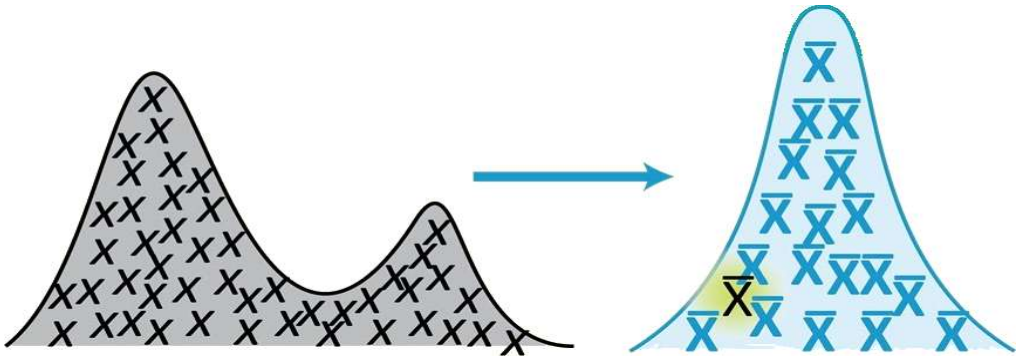
The **sample mean** $\overline{X}_n$ is the first sample moment: $\overline{X}_n := M_1 = \frac{1}{n}\sum_{j=1}^{n} X_j$.

In contrast, the **population mean** or **true mean** is $E(X_j)$, the mean of the distr from which the X_j were drawn.

Law of Large Numbers (LLN): $M_k \to E(X_1^k)$ as the number of rvs goes to infinity.
The sample moment converges to the dist's moment. (more in Chapt. 10)

Lemma (Unbiased Sample Moments). The expected value of the kth sample moment is the kth moment.

Proof. $E\left(\frac{1}{n}\sum_{j=1}^{n} X_j^k\right) = \frac{1}{n}(E(X_1^k) + \ldots + E(X_n^k)) = E(X_1^k)$. ■



PF of iid X_i. (w/1st moment μ) PF of $\frac{1}{n}\Sigma_j X_j$ (w/1st sample moment μ)

Observe the mean μ is the same for both

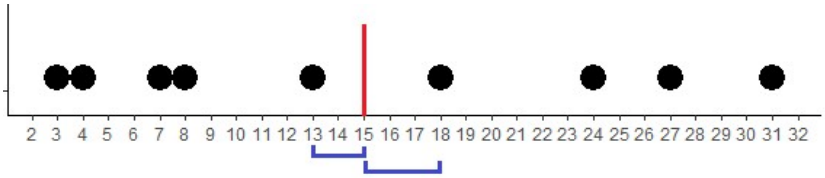
Thm (**Mean and Variance of Sample Mean**): Let $X_1, \ldots, X_n$ be iid w/mean μ and variance σ^2.
Then the sample mean $\overline{X}_n$ is unbiased for estimating μ. That is, $E(\overline{X}_n) = \mu$.
The variance of $\overline{X}_n$ is given by $Var(\overline{X}_n) = \frac{\sigma^2}{n}$.

Proof. We have $E(\overline{X}_n) = \mu$ from the Unbiased Sample Moments lemma.

For variance, we use the fact (shown in chapter 7) that the variance of the sum of indep rvs is the sum of the variances:
$Var(\overline{X}_n) = \frac{1}{n^2} Var(X_1 + \ldots + X_n) = \frac{n}{n^2} Var(X_1) = \frac{\sigma^2}{n}$. ■

Def (**Sample Variance and SD**): Let $X_1, \ldots, X_n$ be iid. The sample variance
is the rv: $S_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2$. (i.e., the average squared distance to the mean)

The sample SD is the square root of the sample variance.



Calculating Variance

Thm (**Unbiasedness of Sample Variance**). Let $X_1, \ldots, X_n$ be iid w/mean μ and variance σ^2.
Then the sample variance S_n^2 is unbiased for estimating σ^2, i.e., $E(S_n^2) = \sigma^2$.

Proof. The key to the proof is the handy identity: $\sum_{j=1}^n (X_j - c)^2 = \sum_{j=1}^n (X_j - \overline{X}_n)^2 + n(\overline{X}_n - c)^2$
which holds for all c. To verify the identity, adding subtract $\overline{X}_n$ in the left-hand sum;

$$\begin{aligned}\sum_{j=1}^n (X_j - c)^2 &= \sum_{j=1}^n ((X_j - \overline{X}_n) + (\overline{X}_n - c))^2 \\ &= \sum_{j=1}^n (X_j - \overline{X}_n)^2 + 2\sum_{j=1}^n (X_j - \overline{X}_n)(\overline{X}_n - c) + \sum_{j=1}^n (\overline{X}_n - c)^2 \\ &= \sum_{j=1}^n (X_j - \overline{X}_n)^2 + n(\overline{X}_n - c)^2.\end{aligned}$$

For the last line, we use the fact that $\overline{X}_n - c$ does not depend on j and the fact that:

$$\sum_{j=1}^n (X_j - \overline{X}_n) = \sum_{j=1}^n X_j - \sum_{j=1}^n \overline{X}_n = n\overline{X}_n - n\overline{X}_n = 0.$$

Now let us apply the identity, choosing $c = \mu$. Taking the expectation of both sides,

$$nE(X_1 - \mu)^2 = E\left(\sum_{j=1}^n (X_j - \overline{X}_n)^2\right) + nE(\overline{X}_n - \mu)^2.$$

By definition a variance, $E(X_1 - \mu)^2 = Var(X_1) = \sigma^2$, and $E(\overline{X}_n - \mu)^2 = Var(\overline{X}_n) = \frac{\sigma^2}{n}$.

Plugging these result in above & simplifying, we have: $E(S_n^2) = \sigma^2$. ■

Similarly, we can define the **sample skewness** to be: $\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^3}{S_n^3}$,

and the **sample kurtosis** to be: $\frac{\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^4}{S_n^4} - 3$.

What did we learn?

- Median/Mode
- Moment/Central/Standardized
- Skewness/Kurtosis
- Sample Moments/Law of Large Numbers