3 Moments & Deviations

3.1 Markov's Inequality

Slide 1

Markov's Inequality

Some useful bounds for tail probabilities, which are useful, for example, in analyzing algorithms.
We start with an easy to prove, but also very weak bound:

Theorem 3.1: Markov's Inequality. Let $X \geq 0$ be a r.v. Then for any $a > 0$

$$\Pr(X \geq a) \leq \frac{E(X)}{a}$$

Proof: Let $I = \mathbb{1}(X \geq a)$. Then by construction (i) $I \leq X/a$, and (ii) $E(I) = \Pr(I = 1) = \Pr(X \geq a)$.

$$\implies E(I) = \Pr(X \geq a) \leq E(X/a) = E(X)/a.$$

Example: $X = \#$ heads in n fair coin flips, $X \sim \mathrm{Bin}(n, \frac{1}{2})$.

$$\implies \Pr(X \geq \frac{3}{4}n) \leq \frac{E(X)}{\frac{3}{4}n} = \frac{\frac{n}{2}}{\frac{3}{4}n} = \frac{2}{3}$$

3.2 Variance & Moments

Slide 2 The first moment is the expected value

Variance & Moments

We can get better bounds by using $E(X^k)$:

Definition 3.1: $E(X^k) = k$-th moment of a r.v. X.

Definition 3.2: $\mathrm{Var}(X) = E\{(X - EX)^2\} = \ldots = E(X^2) - (EX)^2$.

And *standard deviation* $\sigma(X) = \sqrt{\mathrm{Var}(X)}$, or just σ (if the r.v. X is understood)

Proof: of the identity for $\mathrm{Var}(X)$: E(X) = ∑x*Px(x)

$$\begin{aligned} E\{(X - EX)^2\} &= E\left\{X^2 - 2X \cdot EX + (EX)^2\right\} = \\ &= E(X^2) - 2EX \cdot E(X) + (EX)^2 = E(X^2) - (EX)^2 \end{aligned}$$

3.3 Covariance

Slide 3 Positive correlation, Cov(X,Y)>0
Negative correlation, Cov(X,Y)<0

Covariance

While $E(X + Y) = E(X) + E(Y)$, the same is not true for $\mathrm{Var}(X)$.

We need:

Cov(X, Y) = ∑(X – EX)(Y – EY)P(X,Y)

Definition 3.3: Covariance. $\mathrm{Cov}(X, Y) = E\{(X - EX)(Y - EY)\}$

Theorem 3.2. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

Proof: $\mathrm{Var}(X + Y) =$

$$\begin{aligned} & E\left\{[(X + Y) - E(X + Y)]^2\right\} \\ = \; & E\left\{[X + Y - EX - EY)]^2\right\} \\ = \; & E\left\{[(X - EX) + (Y - EY)]^2\right\} \\ = \; & E\{(X - EX)^2\} + E\{(Y - EY)^2\} + 2\,E\,\{(X - EX)(Y - EY)\} \\ = \; & \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) \end{aligned}$$

Slide 4

More on (Co-)Variance

Theorem 3.3: If X, Y are independent r.v.'s (we write "$X \perp Y$"), then

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

but, the same is *not true for dependent r.v.'s.*

Proof: $E(XY) = \sum_i \sum_j (i \cdot j) \underbrace{\Pr(X = i, Y = j)}_{= p_X(i) \cdot p_Y(j)} =$

$$\left(\sum_i i p_X(i)\right) \cdot \left(\sum_j j p_Y(j)\right) = E(X) \cdot E(Y)$$

Corollary 3.4: $X \perp Y \implies \mathrm{Cov}(X, Y) = 0$ and therefore

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Independent, joint probability function = product of the marginal probability functions

Proof: $\mathrm{Cov}(X, Y) =$

$$= E\{(X - EX) \cdot (Y - EY)\} = E(XY) - EX \cdot EY - EX \cdot EY + EX \cdot EY = 0$$

using $E(XY) = EX \cdot EY$.

3.4 Example: Binomial Variance

Slide 5

Example: Binomial Variance

Recall binomial r.v.'s. Let $X \sim \mathrm{Bin}(n, p)$, with

$$p_X(j) = \binom{n}{j} p^j (1 - p)^{(n-j)}.$$

We find $\mathrm{Var}(X)$. Some observations

- Binomial theorem:

$$\sum_{\ell=0}^{k} \binom{k}{\ell} p^\ell (1 - p)^{(k-\ell)} = 1.$$

- We will find $\mathrm{Var}(X) = E(X^2) - (EX)^2$. In the expansion we will twice use the binomial theorem, with $k = n - 2$ and $k = n - 1$, respectively.

Slide 6

$$\begin{aligned}
E(X^2) &= \sum_j p_X(j)\cdot j^2 = \sum_{j=0}^{n}\binom{n}{j}p^j(1-p)^{n-j}\cdot((\overbrace{j^2-j}^{j(j-1)})+j)\\
&= \sum_{j=0}^{n}\frac{n!(j^2-j)}{(n-j)!\,j!}p^j(1-p)^{n-j}+\sum_{j=0}^{n}\frac{n!}{(n-j)!\,j!}p^j(1-p)^{n-j}\cdot j
\end{aligned}$$

n: n-2, k: n-j, n-k: (n-2)-(n-j)=j-2

$$\begin{aligned}
&= n(n-1)p^2\underbrace{\sum_{j=2}^{n}\frac{(n-2)!}{(n-j)!(j-2)!}p^{j-2}(1-p)^{n-j}}_{=1 \text{ (bin. thm., } k=n-2,\ \ell=j-2)}+\\
&\quad np\underbrace{\sum_{j=1}^{n}\frac{(n-1)!}{(n-j)!(j-1)!}p^{j-1}(1-p)^{n-j}}_{=1 \text{ (bin. thm., } k=n-1,\ \ell=j-1)}\\
&= n(n-1)p^2+np
\end{aligned}$$

j! = j(j-1)(j-2)!

j! = j(j-1)!

and therefore $\mathrm{Var}(X) = E(X^2) - (EX)^2 = \ldots = np(1-p)$. E(X) = np

n choose k = n!/k!(n–k)!

3.5 Examples

Slide 7

Example

Suppose that it is known that the number of items X produced in a factory during a week is a random variable with mean $\mu = 50$.

(a) What can be said about $\Pr(X \geq 75)$?

Solution: Markov's inequality gives $\Pr(X \geq 75) \leq \mu/75 = 2/3$. [1]

(b) If $\mathrm{Var}(X) = \sigma^2 = 25$, then what can be said about $\Pr(40 < X < 60)$?

Solution: Consider $D = (X-\mu)^2$.
Note $\mathrm{Var}(X) = E[(X-\mu)^2] = E(D)$.

By Markov inequality (for $D \ldots$)

Pr(40<X<60) = Pr((X-µ)<10)

$$1 - \Pr(40 < X < 60) = \Pr(D > 10^2) \leq \frac{E(D)}{100} = \frac{\sigma^2}{100} = \frac{1}{4}.$$

Pr((X-µ)>10)

[2] (Note: see Chebyshev's inequality).

[1] In the lecture video we used $X > 75$. Considering the integer nature of X (as a count) we could actually get a sharper bound by $\Pr(X > 75) = \Pr(X \geq 76) \leq \mu/76$.

[2] In the lecture the previous line was missing $1 - \ldots$!

3.6 Chebyshev's Inequality

Slide 8

Chebyshev's Inequality

Recall Markov's inequality, using only $E(X)$. Using also the 2nd moment, $\mathrm{Var}(X)$ we get a better bound

Theorem 3.6: Chebyshev's Inequlity. If X is a r.v. with $\mathrm{Var}(X)$, then for any $a > 0$,

$$\Pr(|X - EX| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

Var(X) = E{(X – EX)^2}

Proof: Markov inequality with $D = (X - EX)^2$.

Corollary 3.7. For any $t > 1$

$$\Pr(|X - EX| \geq t\sigma(X)) \leq \frac{1}{t^2}$$

and

$$\Pr(|X - EX| \geq tEX) \leq \frac{\mathrm{Var}(X)}{t^2(EX)^2}$$

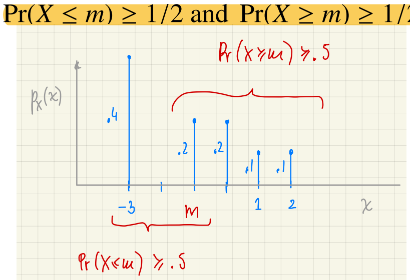
3.7 Median & Mean

Slide 9

Median and Mean

Median: the *median* of a r.v. X, $\mathrm{Md}(X)$, is any value m with

$$\Pr(X \leq m) \geq 1/2 \text{ and } \Pr(X \geq m) \geq 1/2$$



Example: $X \sim \mathrm{Unif}(\{x_1, \ldots, x_{2k+1}\})$, uniform over an odd # of (sorted) values. Then $\mathrm{Md}(X) = x_{k+1}$.

If $Y \sim \mathrm{Unif}(\{x_1, \ldots, x_{2k}\})$, then $\mathrm{Md}(Y) = m$ for any value $x_k < m < x_{k+1}$.

Slide 10 E(X) = ∑x*Px(x)

Theorem 3.9: For any r.v. X with finite $E(X)$ and $\mathrm{Md}(X)$

(a) $E(X) = \mu$ is the value c that minimizes $E[(X-c)^2]$, and

(b) $\mathrm{Md}(X) = m$ is *a* value c that minimizes $E[|X-c|]$ (need not be unique).

Proof: Part (a) follows from taking the derivative in $E[(X-c)^2] = E(X^2) - 2cEX + c^2$, giving

$$-2\,EX + 2c = 0 \implies c = EX.$$

when derivative = 0, get minimizes

Slide 11

For (b), will show $E(|X-c| - |X-m|) \geq 0$ for any $c \neq m$.
First consider a value c with $c > m$ and therefore $\Pr(X \geq c) < 1/2$
(and then a similar argument for $c < m$),

Write the expectation, breaking it down into three terms:

$$|x-c| - |x-m| = \begin{cases} -(c-m) & \text{for } x \geq c \\ c+m-2x & \text{for } m < x < c \\ c-m & \text{for } x \leq m, \end{cases}$$

implying

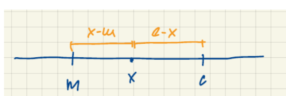
$$E(|X-c| - |X-m|) = -(c-m)\,\Pr(X \geq c) + \\ + \sum_{m<x<c} (c+m-2x)\,p_X(x) + (c-m)\,\Pr(X \leq m)$$

Case $\Pr(m < X < c) = 0$**:** We are done:

$$|x-c| - |x-m| = -(c-m)\,\Pr(X \geq c) + (c-m)\,\Pr(X < c) > 0$$

recall that we are in the case $c > m$ and $\Pr(X \geq c) < 1/2$.

Slide 12



Case $\Pr(m < X < c) > 0$**:**
$E(|X-c| - |X-m|) =$

$$\begin{aligned} &= -(c-m)\,\Pr(X \geq c) \\ &\quad + \sum_{m<x<c} \underbrace{((c-x) - (x-m))}_{\geq -(c-m)}\, p_X(x) + (c-m)\,\Pr(X \leq m) \\ &> -(c-m)\,\underbrace{\Pr(X > m)}_{<0.5} \qquad + (c-m)\,\underbrace{\Pr(X \leq m)}_{\geq 0.5} \\ &> -(c-m)\,0.5 + (c-m)\,0.5 = 0. \end{aligned}$$

In the 2nd line we used $(c-x) - (x-m) > -(c-m)$: the difference of the two subintervals $> -$ length of the entire interval.

Slide 13

Theorem 3.10: If X is a r.v. with finite $E(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ and $\mathrm{Md}(X) = m$, then $|\mu - m| < \sigma$. Var(X) = E{(X – EX)2}

Proof:

$$\begin{aligned} |EX - m| &= |E(X-m)| \leq E(|X-m|) \leq \\ &\leq E(|X-\mu|) = E[\sqrt{(X-\mu)^2}] \leq \sqrt{E[(X-\mu)^2]} = \sigma \end{aligned}$$

Md(X) = m is a value c that minimizes E[|X – c|]

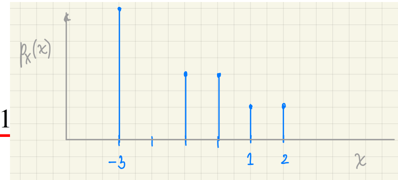
using Jensen's inequality (1st and 3rd inequality) and Theorem 3.9(b) (2nd inequality).

3.8 Examples

Slide 14

Examples

Consider a discrete r.v. X with



$$p_X(x) = \begin{cases} 0.4 & \text{for } x = -3 \\ 0.2 & \text{for } x = -1 \\ 0.2 & \text{for } x = 0 \\ 0.1 & \text{for } x = 1 \\ 0.1 & \text{for } x = 2 \end{cases}$$

(a) Find a to minimize the expected sq distance $E((a-X)^2)$.

Solution: By Theorem 3.9a, $a = E(X) = -1.1$.

(b) Find b to minimize the expected absolute distance $E(|b-X|)$.

Solution: By Theorem 3.9b, $b = \mathrm{Md}(X) = -1$.

Slide 15

For those following along in the book, we are now skipping §3.5.

We are now ready to discuss statistical inference, using, for example, binomial r.v.'s and the rules for working with probabilities.

3.9 The Role of Probability in Statistical Inference

Slide 16

The Role of Probability in Statistical Inference

Probability: In the discussion so far, did you notice that we cheated? We always assumed that probabilities and distributions of r.v.'s were *known*.

For example "Let X be the number of cells with [. . .]. Assume $X \sim \mathrm{Bin}(n, p)$, with $p = 0.6$. Find $\Pr(X \geq 100)$."

In a real experiment, e.g., carried out in a lab, you *never know* p! Or the rate λ of the service time for a customer in the post office etc. This is where statistics comes in.

Statistics: We ask "Using an observed value x for $X \sim \mathrm{Bin}(n, p)$, can we guess what p could be?", or "Could p be greater 0.5?"

That is, we change perspective. Assuming we have a good description of the experimental data as a r.v.'s, we try to report inference on the parameters, like p etc.

Slide 17

The change of perspective between probability and statistics naturally gives rise to some different vocabulary in statistics.

- *Sample*: a set of r.v.'s $\boldsymbol{X} = \{X_i, i = 1, \ldots, n\}$, $X_i \sim F$, i.i.d. that are actually observed. We also just refer to $\boldsymbol{X}$ as the *data*. set of random variables

- *Statistic*: any function of the observed data, $S = f(X_1, \ldots, X_n)$. In probability we would have simply said S is another r.v. Of course, it is both!

- *Parameters:* This is tricky. In some pars of statistics (Bayesian statistics), parameters are r.v.'s that are not observed.
  In other contexts they index (describe) the distribution of the data. We usually use greek letters, like λ, μ, well, or p.

- *Hypothesis:* a hypothesis is simply an event for the parameters, like $A = \{\mu > 0\}$ (and we might use different names like H_0 etc.)

In summary, probability describes uncertainty, whereas statistics is about decisions in the face of such uncertainty.