# Toward Statistical Inference

*Thanks to Joe Chang in the Statistics Department for some lecture content*

**Inference** – use information about a sample to draw an inference about the population

*I'm starting to get concerned about Global Warming!*

***Example** : A Gallup poll of 1000 people reveal that in 2006, about 80% of Americans believed that Global Warming was actually happening (compare to 22% in 1991 – note that a Gallup poll last March shows that 49% believe the severity is exaggerated (67% for conservatives). We turn the <u>fact</u> that 80% of the sample have this opinion into an <u>estimate</u> that 80% of all Americans feel this way.*

Remember :

**Parameter** – a fixed number that describes a **population** (i.e., $\mu$ = true population mean height). We don't know this number (Gods only)

**Statistic** – a number that describes a **sample** (i.e., $\bar{x}$ = sample mean height). We know this number, <u>but</u> the number can (and usually does) **change from sample to sample**. Use the statistic to estimate the unknown parameter!



**Sampling Variability** – If we repeated our sampling procedure 'many' times, the same way each time, how much would our statistics change from one sample to the next?

**Sampling Distribution of a statistic** – the distribution of values of a sample statistic in all possible samples of the same size from a fixed population.

***Example** : Let $p$ be the **true proportion** of the population that believes in global warming (the **PARAMETER).***

*Suppose a <u>**TOTAL of FOUR**</u> people live in the U.S. (this is the **population)**. I.e., just this once, we know the entire population. Here are their opinions (known to Gods, who are letting us know just this once . . . )*

| ***Individual*** | ***Attitude*** |
|---|---|
| *1* | *Believe* |
| *2* | *Believe* |
| *3* | *Don't Believe* |
| *4* | *Don't Believe* |

- *We want to estimate $p$ using a **STATISTIC** $\hat{p}$, the sample proportion that believes in global warming.*

*In this population, $p$=0.5 (the **parameter,** i.e. the **true** proportion of the population that believes in global warming).*

***NOW :*** *Pretend we don't know $p$=0.5, so we take a sample of size $n$=2.*

*List **all possible** Simple Random Samples (SRS) of size 2 from this population, and record the sample proportion for each sample (the **statistic**, $\hat{p}$)*

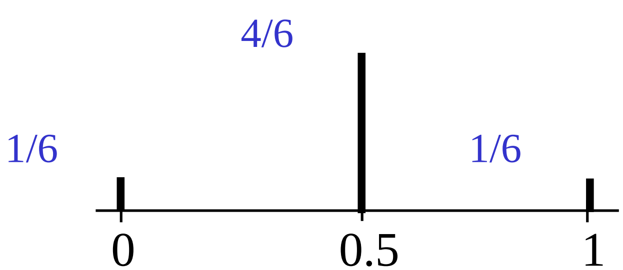
***POPULATION***

| ***Individual*** | ***Attitude*** |
|---|---|
| *1* | *Believe* |
| *2* | *Believe* |
| *3* | *Don't Believe* |
| *4* | *Don't Believe* |

***POSSIBLE SAMPLES***

| ***Individuals in SRS*** | ***Attitude*** | $\hat{p}$ |
|---|---|---|
| *1 2* | *B B* | 1 |
| *1 3* | *B DB* | 0.5 |
| *1 4* | *B DB* | 0.5 |
| *2 3* | *B DB* | 0.5 |
| *2 4* | *B DB* | 0.5 |
| *3 4* | *DB DB* | 0 |

*In terms of probability, this is the **sampling distribution** for $\hat{p}$ for samples of size two :*



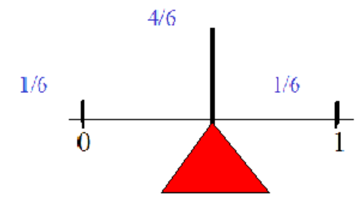

The sampling distribution gives



- All possible values of $\hat{p}$
- The **proportion** of times $\hat{p}$ takes on each of these values, or the **probability** that $\hat{p}$ takes each of these values.

**Now :** what is the average value of the sample proportion, $\hat{p}$?. This is called the **Mean of a sampling distribution.**

**Mean of a sampling distribution = balancing point**




***In this case, the*** *Mean of sampling distribution of $\hat{p}$ = 0.5*

*This is also the value of the parameter $p$, the <u>true</u> proportion of the population that believes in global warming.*

| |
|---|
| If the mean of the sampling distribution of a statistic equals the true value of a parameter, the statistic is said to be an **UNBIASED ESTIMATOR** of the parameter |

**Now :** what is the variability of the sampling distribution of $\hat{p}$?

*We'll see formulas for calculating this later. Suffice it to say that the standard deviation of $\hat{p}$ is about 0.32. Trust me.*

---

***NOW*** *: Suppose our budget is so small we can only afford a sample of size $n$=1. How does the sampling distribution of $\hat{p}$ change?*

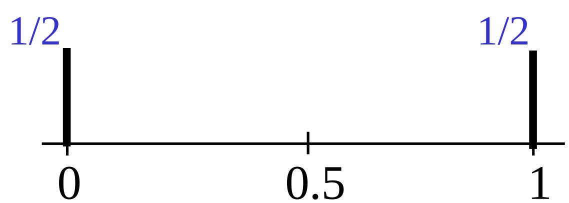
***POPULATION***

| ***Individual*** | ***Attitude*** |
|---|---|
| *1* | *Believe* |
| *2* | *Believe* |
| *3* | *Don't Believe* |
| *4* | *Don't Believe* |

***POSSIBLE SAMPLES***

| ***Individuals in SRS*** | ***Attitude*** | $\hat{p}$ |
|---|---|---|
| 1 | *Believe* | 1 |
| 2 | *Believe* | 1 |
| 3 | *Don't Believe* | 0 |
| 4 | *Don't Believe* | 0 |

*Sampling Distribution of $\hat{p}$ for $n$=1 :*






*$\hat{p}$ is still* ***unbiased****: Mean of sampling distribution = 0.5 = true proportion that believe in global warming, $p$*

***Variability*** *of the sampling distribution of $\hat{p}$ is clearly 0.5 in this case.*

---

***SO :*** *Mean of sampling distribution of $\hat{p}$ = .5 for samples of size 1 or 2.*

***However :***

***Standard Dev.*** *of $\hat{p}$ with samples of size 2 ≈ 0.3*
***Standard Dev.*** *of $\hat{p}$ with samples of size 1 ≈ 0.5*

*Taking samples of size 2 seems to give us estimates that are* ***less variable!***

# BIAS and VARIABILITY

**Bias of an estimator**

= (mean of sampling distrib.) - (true value of parameter)

Statistic is **unbiased** if bias = 0.

Taking a random sample guarantees that a statistic will be unbiased.

**Variability** of an estimator
= (Standard Deviation of sampling distrib.)

Variability is a function of sample size – larger sample size = less variability in estimator

**To see this, we simulate (make up) data. . . . .**

***Example :*** *Suppose the true proportion of people who believe in global warming is 0.80, or 80%.*

- *We now assume that the population is large, and we just happen to know the true proportion of believers (i.e. $p$ = 0.80)*



- *If we take a sample of size $n$=10, what sort of values for $\hat{p}$ (the sample proportion) might we see? How often will be see these particular values? This is the **sampling distribution**.*

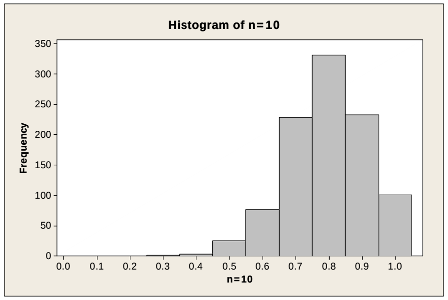
- *To estimate the sampling distribution, let's simulate taking many samples of size $n$=10 (how about 1000 such samples), and make a histogram of how often we see each value of $\hat{p}$ .*

Random Binomial Data in MINITAB : use `Calc → Random Data → Binomial` (more on the binomial distribution later). The number of trials is 10, the probability of success is 0.8, we generate 1000 rows of data.

SPSS: This is what I can figure out : start in excel, make a spreadsheet with the numbers 1 though 1000 in one column. Import into the SPSS. Then use `Transform → Compute.` Use the function `RV.BINOM(10,.8)` (first number is n, second is p, probability)

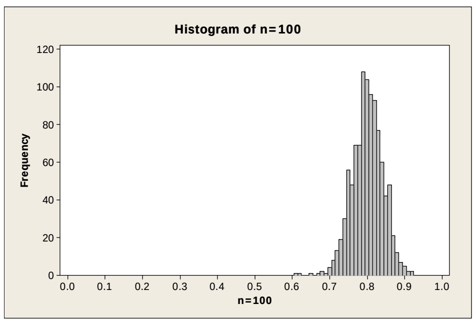
***Note*** *: this is just an **EXERCISE** – you would never actually take many samples of size 10, only **ONE** sample of size 10. We are doing this to see what values (and with what frequency) our sample proportion $\hat{p}$ might take! Another way to think about it : about how far can $\hat{p}$ be from the true value 0.80 just by chance?*

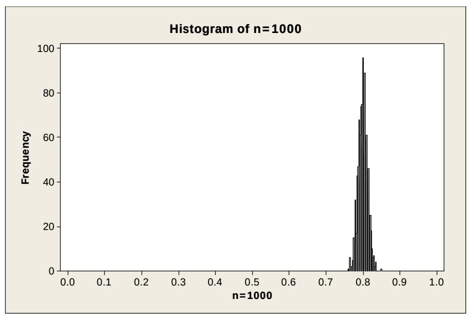
*Here is the estimated sampling distribution of $\hat{p}$ for samples of size $n$ = 10 :*




*Now, suppose we took samples of size $n$=100. How does the estimated sampling distribution of $\hat{p}$ change?*

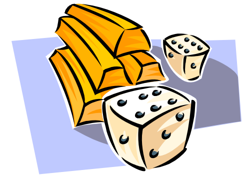
*Now, suppose we took samples of size $n$=1000. How does the estimated sampling distribution of $\hat{p}$ change?*




Clearly, **for larger samples**, the **variability of the sampling distribution decreases** : that is, on average our sample statistic $\hat{p}$ will generally be closer to the true parameter $\hat{p}$ for larger sample sizes.





# PROBABILITY

*Chapter 3 in Cartoon Guide – STRONGLY RECOMMENDED*

**Probability** is crucial to **statistical inference**

- Inferences are always expressed in terms of probability

  *(i.e. a “95% Confidence Interval”. 0.95 is the probability of something . . .)*

A survey – Suppose exactly 80% of people believe in global warming.

- We take a random sample of size 100
- Expect to see about 80 believe in global warming.
- How likely are we to observe more than 90 who believe in global warming?

**Probability Models** – We’re modeling some random phenomenon. A probability model consists of

- A List of possible outcomes
- A probability for each outcome

**Sample space** = $S$ = set of all possible outcomes

***Examples :***

<u>**Discrete**</u>

*Toss a coin:* $S$ *= {H,T}.*

*Watch a tree for a year and see if it dies :*
$S$ *= {Dead, Alive}.*

*Toss a coin 3 times:*

$S$ *= {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}*

<u>**Continuous**</u>

*Record the gas mileage for a car :*
$S$ *= {a positive number between 0 and ? (100mpg)?}*

An **event** : set of *some* possible outcomes :

- An event is a <u>subset</u> of outcomes in $S$
- Denoted by $A$ (or $B$ or $C$)

***Example** : Let the event* $A$ *= (get one head in 3 tosses)*

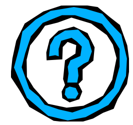
*{HHH, HHT, HTH, $\boxed{\text{HTT,}}$ THH, $\boxed{\text{THT,}}$ $\boxed{\text{TTH,}}$ TTT}*

*= {HTT, THT, TTH}*



**Probability measure** : a function (satisfying certain conditions) that assigns a ***probability*** (a number between 0 and 1) to each event.

If $A$ is an event, $P(A)$ denotes the probability of $A$.



**SO** : What does probability mean, and how do we assign probabilities to outcomes?!? (i.e., how do we define a probability measure?)

**Three approaches :**

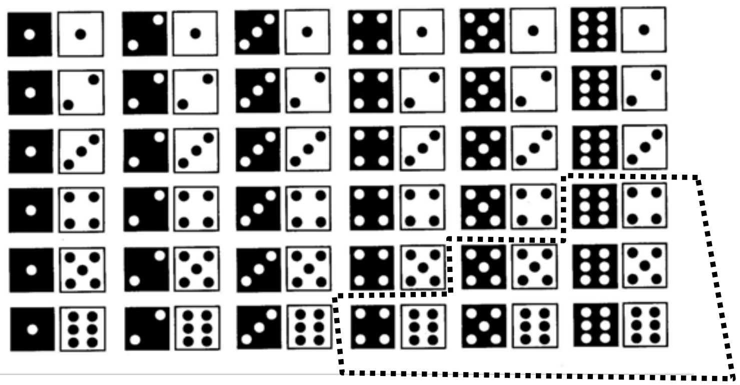
1) *Classical*
2) **Relative Frequency**
3) Personal/Subjective Probability (Baysian)

1) *Classical* (based on gambling): Sometimes, we believe all possible outcomes are equally likely (i.e. the game is fair!). In this case

$$P(A) = \frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S}$$

***Example :** Toss a coin once,* $S=\{H,T\}$.
*If* $A=\{H\}$, then $P(A)=0.5$

***Example :*** *Roll Two Dice* $S=\{\textit{see picture}\}$



*If* $A=\{$*Sum of Dice at least 10*$\}$, then

$$P(A)=\frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S}=\frac{6}{36}=0.167$$

2) **Relative Frequency ( Long Run Frequency ):** When an experiment can be repeated, the probability of an event is the proportion of times the event occurs in the long run


***Example :*** *What's the probability a radish seed will grow in soil treated with RoundUp? Plant many, many seeds, count the number of times the seed germinates.*



3) Personal/Subjective Probability (Baysian) :

Most events in life aren't repeatable. We assign probabilities all the time :

- What's the probability I'll take statistics?
- What's the probability this guy will ask me out on a date?
- What's the probability a huge body of fresh water will halt the gulf stream and lead to an ice age within a century? (some people think high . .)

Think in terms of betting . . . .

This may seem arbitrary, but it's actually a good description of how we assign probabilities. Baysians assign a probability based on the information at hand, **and update their probability as more data becomes available.**

---

Regardless of how you choose to assign probabilities, the following two rules apply :

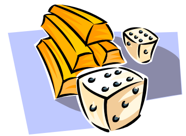
1) For any event $A$, $0 \le P(A) \le 1$

*(Probabilities are always between zero and one)*

2) $P(S)=1$

*(Something must happen)*

***Now a bit of gambling history . . . . .***



A rich Frenchman, Antoine Gombaud, known as the **'Chevalier de mere'** liked to gamble. However, he was confused by certain experiences at the gambling tables. He posed the following question to his mathematician friend, **Blaise Pascal** in 1654 :

*Which is more likely :*

1) *At least one six in four rolls of a single die*

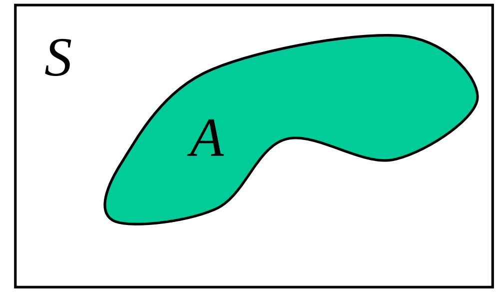
2) *At least one double-six in 24 rolls of a pair of dice*

Pascal with his friend **Pierre de Fermat** soon worked out the Chevalier's problem, and in the process developed the algebraic basis of probability theory.

*Here is the theory they worked out . . .*



**Venn Diagrams** : Represent Sample Spaces and Events with **pictures!**




- Think about $S$ as your car windshield
- $A$ is an area in the windshield.
- It's about to start raining.
- Let $A$ be the event that the first drop lands in the area $A$.
- Rain is equally likely to fall anywhere on the windshield.

Probability measure for this picture : **Probability = Area**

$$P(A) = \frac{\text{area of } A}{\text{area of } S}$$

For convenience assume (area of $S$ ) = 1.

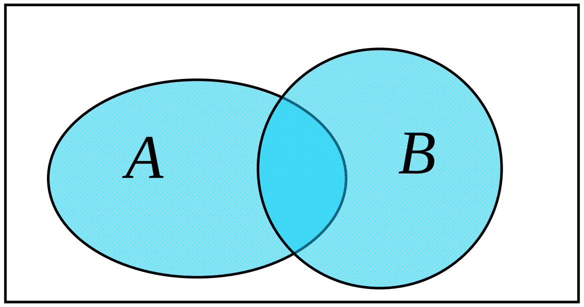
So $P(A) =$ area of $A$.

**Remember** : $0 \leq P(A) \leq 1$ and $P(S) = 1$

## Make New Events from Old Events

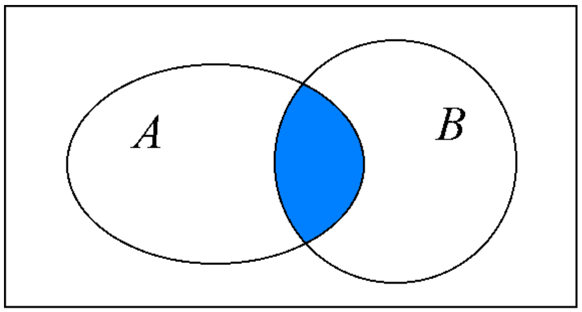
$A$ **or** $B$

(raindrop falls
in A or B)

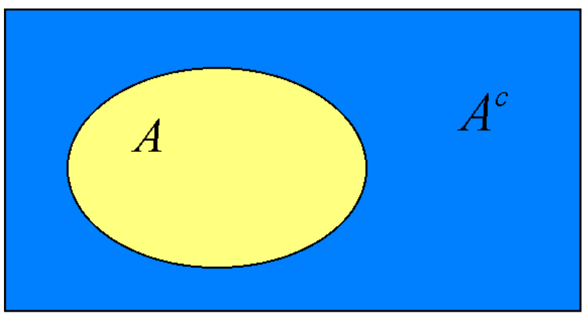



$A$ **and** $B$

(raindrop falls in
A and B)




**Complement** of $A$
(raindrop falls
in 'not A')





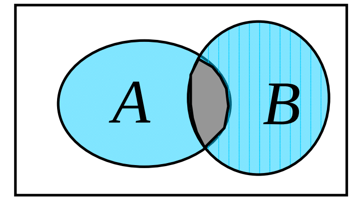

## Axioms of Probability

(properties of Probability Measures)

- For each event $A$, $0 \le P(A) \le 1$
- $P(S) = 1$, where $S$ is the whole sample space.
- **Addition Rule :**

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

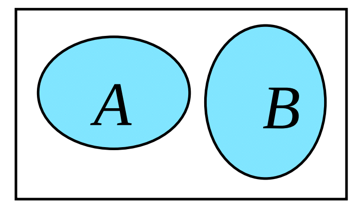
*(Middle shaded area gets counted twice, so we have to subtract this area, which is A and B)*




- **Disjoint Events :** If $A$ and $B$ are disjoint, then

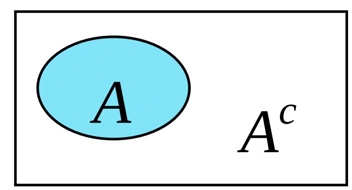
$$P(A \text{ or } B) = P(A) + P(B)$$

i.e. $P(A \text{ and } B) = 0$




- **Complement rule**

$$P(A) = 1 - P(A^c)$$



## Conditional Probability

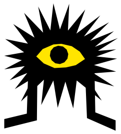
Notation : $P(B \mid A)$

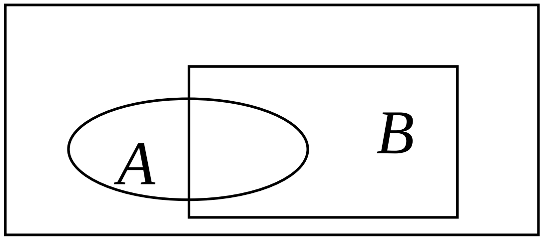
*Read as* " Probability of $B$ given $A$ "

Meaning : Given that $A$ has already happened, what is the probability of $B$?

By eyeball :

$$P(B \mid A) = 0.5$$

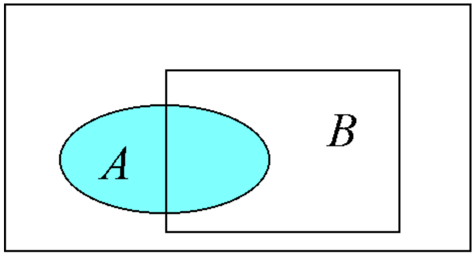




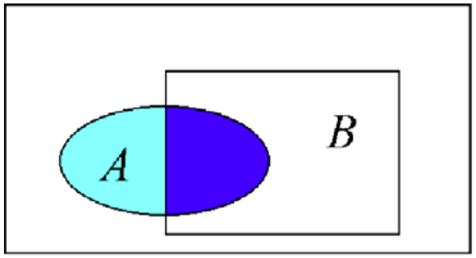



## Conditional Probability (cont.)

Given that the raindrop fell in $A$, we restrict our attention to the set $A$. The drop is equally likely to fall anywhere within $A$.




Given $A$, the event $B$ also occurs when the drop falls in the darker region, i.e., the event ( $A$ and $B$ ).




**This gives the formal definition for Conditional Probability :**

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

**Or, equivalently**

$$P(A \text{ and } B) = P(B \mid A)P(A)$$

## Independence

Two events $A$ and $B$ are **independent** if being told that $A$ occurred has no effect on the probability that $B$ also occurrs.

**Formal Definition :**

$A$ and $B$ are independent if

$$P(B \mid A) = P(B)$$

Equivalently,

$$\frac{P(A \text{ and } B)}{P(A)} = P(B)$$

Equivalently,

$$P(A \text{ and } B) = P(A)P(B)$$

<u>**WARNING :**</u>

**Don't confuse <u>Independent</u> with <u>Disjoint</u>**

Independent Events : $P(B \mid A) = P(B)$

Disjoint Events : $P(A \text{ and } B) = 0$

> **If events are disjoint, they cannot be independent.**
> **If events are independent, they cannot be disjoint.**

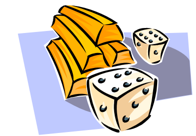


***Example*** *: Toss two coins. Given that the first coin is heads, what is the probability that the second coin is also heads?*

*Coins are independent so*

$$P(\text{Heads (2)} \mid \text{Heads (1)}) = P(\text{Heads (2)})$$

*That is, these events are **independent**.*



***Example*** *: Toss a pair of dice. Let $A$ be the event that the first die equals 5. Let $B$ be the event that the sum of the dice equals 4.*

$$P(A) = 1/6$$
$$P(B|A) = 0$$

So

$$P(A \text{ and } B) = P(A)*P(B|A) = 1/6 *0 = 0$$

***SO :*** *$A$ and $B$ are **disjoint**, but they are **not independent.***

***Back to the Chevalier's problem . . . . .***

*Which is more likely :*

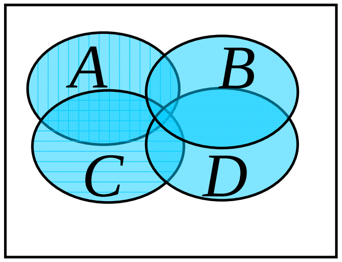
*1) At least one six in four rolls of a single die*

*2) At least one double-six in 24 rolls of a pair of dice*

1) Let the events $A, B, C, D$ be the events of getting a six on roll 1, 2, 3, or 4, respectively.

We want

$$P(A \text{ or } B \text{ or } C \text{ or } D)$$
$$= P(A) + P(B) + P(C) + P(D)$$
$$- \text{ (probability of overlaps)}$$




*Hmm. Overlaps look hard.*

**Better idea : USE COMPLEMENT RULE :**
Get area of region outside the discs.



$$P(\text{at least one six}) = 1 - P(\text{no sixes})$$
$$= 1 - P(A^c \text{ and } B^c \text{ and } C^c \text{ and } D^c)$$
$$= 1 - P(A^c)P(B^c)P(C^c)P(D^c)$$
$$= 1 - \left(\frac{5}{6}\right)^4 = 0.518$$

*Because dice rolls are **independent** – the value of one die throw does not change the likelihood of outcomes on subsequent throws*

*2) At least one double-six in 24 rolls of a pair of dice*

*Similar reasoning (use complement rule), gives that*

$$P(\text{at least one double six}) = 1 - P(\text{no double sixes})$$
$$= 1 - \left(\frac{35}{36}\right)^{24} = 0.491$$

**SO : more likely to get one six in four throws of the dice!**