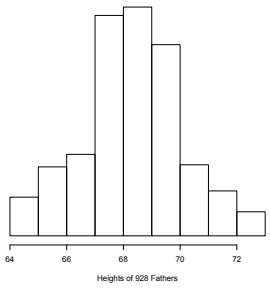
The normal curve

Many data have histograms that look bell-shaped, e.g. heights, weights, IQ scores:



'The data follow the normal curve.'

But remember that some data have histograms that look quite different, e.g. incomes, house prices.

The empirical rule

If the data follow the normal curve, then

- about 2/3 (68%) of the data fall within one standard deviation of the mean
- about 95% fall within 2 standard deviations of the mean
- about 99.7% fall within 3 standard deviations of the mean

Galton's measurements of heights of fathers have $\bar{x} = 68.3$ in and $s = 1.8$ in. Therefore about 95% of all heights are between 68.3 in -2×1.8 in $= 64.7$ in and 68.3 in $+2 \times 1.8$ in $= 71.9$ in.

The empirical rule

Recall that in a histogram, percentages are given by areas:

Standardizing data

A normal curve is determined by $\bar{x}$ and s: If the data follow the normal curve, then knowing $\bar{x}$ and s means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off $\bar{x}$ and then dividing by s:

$$z = \frac{\text{height} - \bar{x}}{s}$$

z is called the **standardized value** or **z-score**.

z has no unit (height, $\bar{x}$ and s all have the unit 'inches')

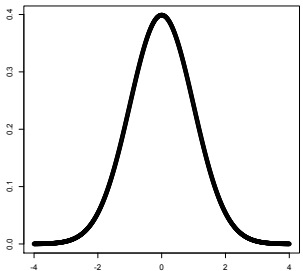
For example, $z = 2$ means the height is 2 standard deviations above average.

$z = -1.5$ means the height is 1.5 standard deviations *below* average.

The standard normal curve

Standardized data have mean 0 and standard deviation equal to 1 – this is the point of standardizing.

Fathers' heights follow the normal curve with $\bar{x} = 68.3$ in and $s = 1.8$ in. Therefore the standardized values follow the **standard normal curve** with mean 0 and standard deviation 1.



This curve is given by the function

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Normal approximation

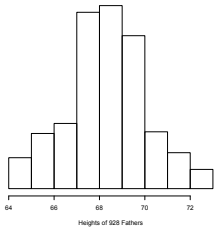
Finding areas under the normal curve is called **normal approximation**.

What percentage of fathers have heights between 67.4 in and 71.9 in?

1. Standardize:

$$\frac{67.4 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = -0.5 \qquad \frac{71.9 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = 2$$

2. Mark the area under the normal curve:



Normal approximation

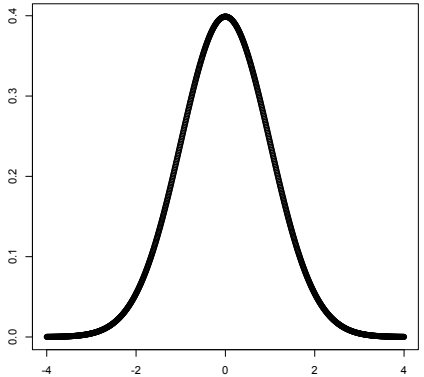
3. Write the desired area in a form that can be computed by software or looked up in a table:

   Typically we can look up the area to the left of a given value.

4. Use software or a table to find these values: $97.7\% - 30.9\% = 66.8\%$

Normal approximation

The empirical rule is a special case of normal approximation:



Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$.

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Or: $z = -0.52$ means that the height is 0.52 standard deviations below average.

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Or: $z = -0.52$ means that the height is 0.52 standard deviations below average. So the height is $\bar{x} - 0.52s = 68.3$ in $-(0.52)(1.8 \text{ in}) = 67.4$ in.

The binomial setting

We saw that for a newborn baby, there is a 49% chance that it is a girl.

What are the chances that 2 out of 3 newborns are girls?

We can compute this by listing all the possibilities (*total enumeration*):

$$\begin{aligned}
\text{P(2 out of 3 are girls)} &= \text{P(GGB or GBG or BGG)} \\
&= \text{P(GGB)} + \text{P(GBG)} + \text{P(BGG)} && \text{addition rule} \\
&= \text{P(G) P(G) P(B)} + \text{P(G) P(B) P(G)} + \ldots && \text{multiplication rule} \\
&= 3 \times (0.49)(0.49)(0.51)
\end{aligned}$$

'3' counts the number of ways one can arrange two G and one B

The binomial setting

This is an example of the **binomial setting**:

- There are $n = 3$ *independent* repetitions of an experiment.
- Each of these experiments has two possible outcomes (which are generically called 'success' and 'failure').
- The probability of success $p = 49\%$ is the same in each experiment.

The binomial coefficient

What about P(2 out of 5 newborns are girls) ?

In principle, we can compute this in the same way. However, there are now 10 possibilities to arrange 2 girls among 5 newborns:

GGBBB, GBGBB, GBBGB,...

The number of possibilities grows very quickly as n gets larger. Fortunately, there is a formula for it:

The **binomial coefficient** counts the number of ways one can arrange k successes in n experiments:

$$\frac{n!}{k!(n-k)!} \qquad \text{where } n! = 1 \times 2 \times 3 \times \ldots \times n$$

$0! = 1$. We had $n = 3$ and $k = 2$, so $\frac{3!}{2!1!} = \frac{1\times2\times3}{1\times2\times1} = 3$.

The binomial formula

Applying this coefficient in the binomial setting gives:

P(k successes in n experiments) $= \frac{n!}{k!(n-k)!}\ p^k(1-p)^{n-k}$

This is the **binomial probability**.

The binomial formula

You play an online game 10 times. Each time there are three possible outcomes: P(win a big prize) $= 10\%$, P(win a small prize) $= 20\%$, P(win nothing) $= 70\%$.

What is P(win two small prizes) ?

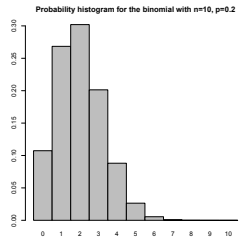
Random variables

The outcomes of the 10 experiments are due to chance, so the number of successes is random: One set of 10 experiments might result in 4 successes, another set might result in 7 successes.

$X =$'number of successes' is called a **random variable**.

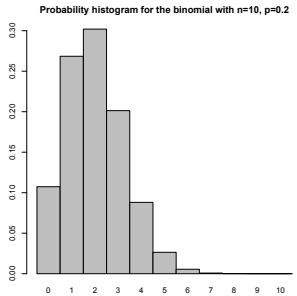
$\mathrm{P}(X = 2) = 30.2\%$. X has the **binomial distribution**.

We can visualize the probabilities of the various outcomes of X with a **probability histogram**:



The probability histogram

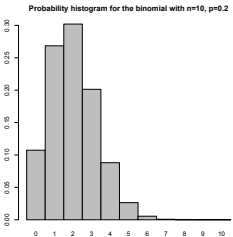
We can visualize the probabilities of the various outcomes of X with a **probability histogram**:

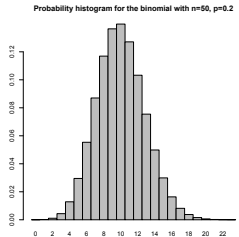


A histogram of data gives percentages for observed data. In contrast, a probability histogram is a theoretical construct: it visualizes probabilities rather than data that have been empirically observed.

Normal approximation to the binomial

As the number of experiments n gets larger, the probability histogram of the binomial distribution looks more and more similar to the normal curve:





In fact, we can approximate binomial probabilities using **normal approximation**: to standardize, subtract off np and then divide by $\sqrt{np(1-p)}$.

Normal approximation to the binomial

We can approximate binomial probabilities using **normal approximation**: to standardize, subtract off np and then divide by $\sqrt{np(1-p)}$.

In the previous example, we had $p =$ P(win a small prize) $= 0.2$.
Play $n = 50$ times. What is P(at most 12 small prizes) ?

Sampling without replacement

A **simple random sample** selects subjects without replacement.

This is not the binomial setting, because p changes after a subject has been removed.

But if the population is much larger than the sample, then sampling with replacement is about the same as sampling without replacement.

Then the number of successes will have approximately the binomial distribution (and so it will approximately follow the normal curve.)