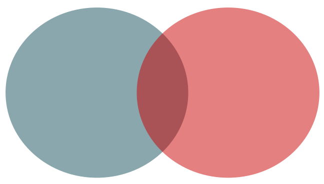
Probabilities and Conditions

Mutually Exclusive events cannot happen at the same time, hence they have no intersect: $P(A \cap B) = 0$

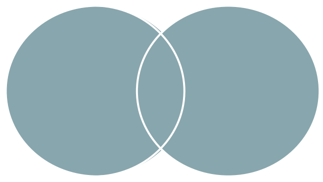
Independent Events have no effect on each other's probabilities: $P(A \cap B) = P(A) \times P(B)$

Cumulative Probability is the sum of probabilities until or beyond a value. They take the form: $P(X > x) = n$ or $P(X < x) = n$

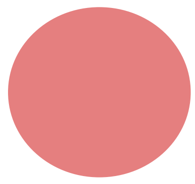
All probabilities add up to 1 or 100% : $\sum P(X = x) = 1$
The most probable, expected value: $\mu = E(X) = \sum x_i \cdot P(X = x)$



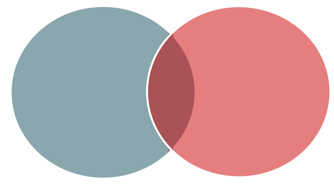
$A \cap B = A \cdot B$ = A intersect B



$A \cup B = A + B$ = A union A



$\bar{A} = A' = 1 - A$ = not A



$A \mid B = (A \cap B) \div B$ = A given B

Sampling and Data Collection Methods

Qualitative Data: Descriptive data about a sample/ population

Quantitative Data: Numerical data about a sample/ population

Discrete Data: Quantitative data that are countable, in natural number values

Continuous Data: Quantitative data that are measured, in real number values

Population: The whole set of things we are interested in

Sample: A subset of the population

Sampling Frame: A list of all members in a population

Population Parameter: A numerical descriptor for a population

Sample Statistic: A value computed from the data of a sample

Census: Data is collected from all members of the population.

- Results are the most accurate it could be
- Time consuming and expensive
- Not feasible for consumable populations (taking a sample from all water bottles produced)

Sampling: Data is collected from a subset of the population.

- Faster data collection and cheaper process
- Less data to analyse
- May not accurately represent the population
- May introduce researcher's bias

Simple Random Sampling: A simple function is used to generate a subset of n members from the larger population

Systematic Sampling: A subset of n members in a population of size N is selected by listing all members and selecting every kth member where $k = N/n$

Stratified Sampling: The population is categorised into strata and a proportionally representative sample is selected randomly from each stratum.

Quota Sampling: The population is stratified and a proportional quota is determined for each strata. However, participation is then voluntary for the population and members are selected until the quota is met

Opportunity Sampling: The sample is selected from the most available members of the defined population

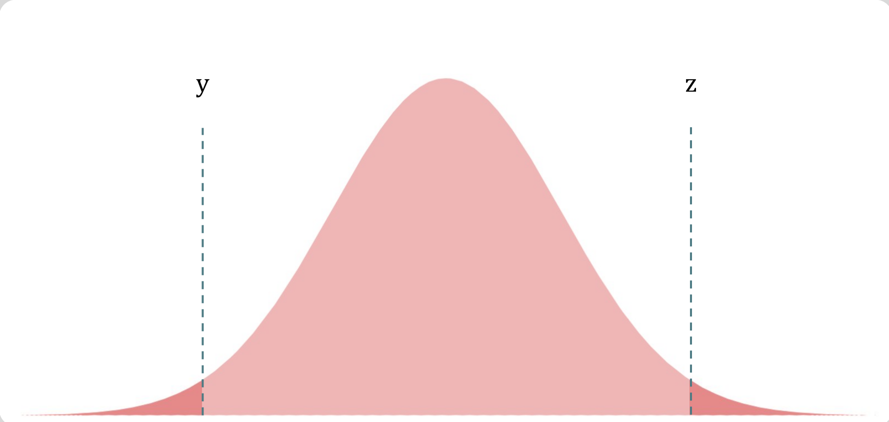
Hypothesis Testing with Distributions and Correlation

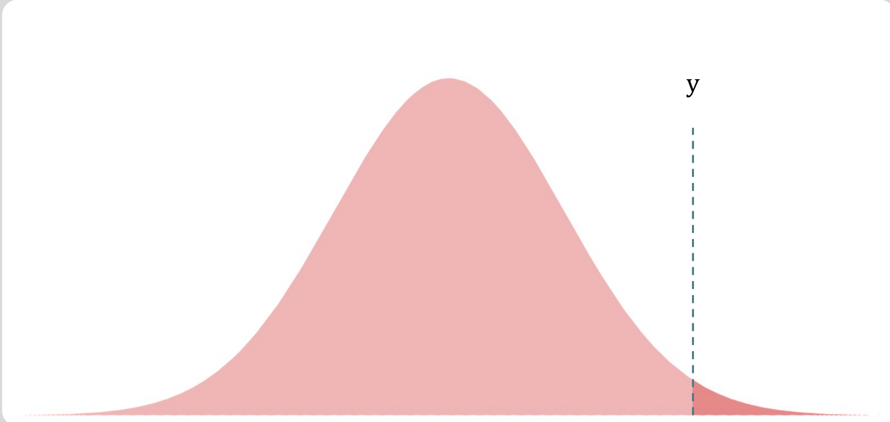
Probability Parameter is the numerical characteristic that defines the population

Significance Level is how unlikely is an event occurring at a given probability. If $P < \alpha$ the event is too unlikely to occur

Critical Region is the set of values where $P < \alpha$ and the range of events are too unlikely to occur

Actual Significance Level is the sum of the probability that the Critical Region(s) may occur





Calculating Critical Region

** The probability may change from α or ½α if this is a Binomial Distribution

$\alpha = x\%$

$P(X < y) = ½\alpha$

$P(X > z) = ½\alpha$

Critical Region: $(x < y)$ and $(x > z)$

Calculating Actual Significance Level

** This is only for Binomial Dist. as the critical boundaries may not be discrete sometimes

Actual Significance = $P(X < y) + P(X > z)$

Calculating Critical Region

** The probability may change from α or ½α if this is a Binomial Distribution

$\alpha = x\%$

$P(X < y) = \alpha$ or,

$P(X > y) = \alpha$

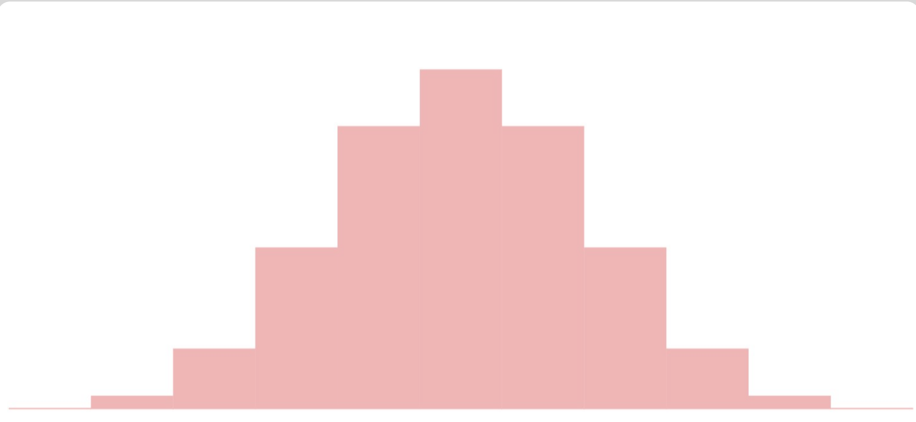
Critical Region is the y inequality for whichever applicable

Calculating Actual Significance Level

** This is only for Binomial Dist. as the critical boundaries may not be discrete sometimes

Actual Significance = $P(X \geqq y)$

Binomial Distribution $X \sim B(n, p)$



This distribution is used with the following requirements:

- There is a fixed n number of trials
- There are only two outcomes (success or fail)
- The probability, p, remains constant
- Each trial is independent from each other

H_0: P(leading digit = 1) = 0.25 (clearly define your P value)
H_1: P(leading digit = 1) > 0.25
$\alpha = 0.05$ (use ½α in two-tailed tests)

Method 1:

$X \sim B(18, 0.25)$
$P(X \geq 8) = 0.056947\ldots \approx 0.0569$
$0.0569 > \alpha$ (if calculated $p < \alpha$ the event is <u>unlikely</u>)

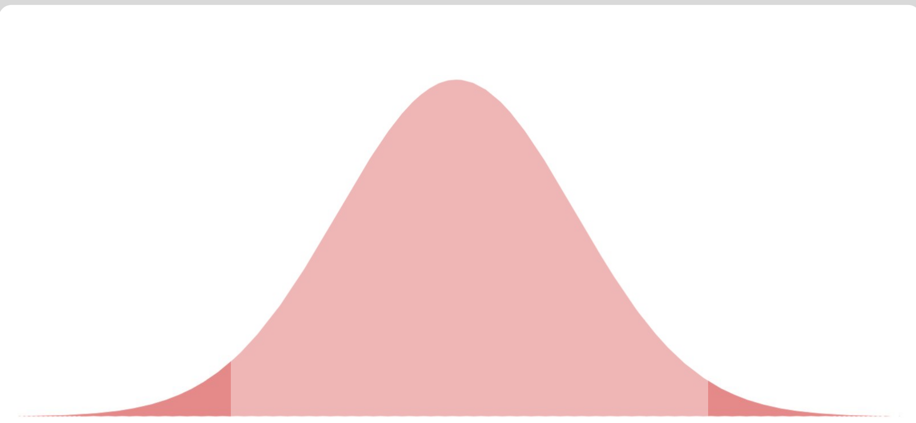
Method 2:

$P(X \geq x) = 0.05$
$\Rightarrow P(X \geq 8) = 0.056947\ldots$
$\Rightarrow P(X \geq 9) = 0.019347\ldots$ (the critical range is ≥ 9)
$9 > 8$ (if given no. of successes are in the critical range, event is <u>unlikely</u>)

Therefore, do not reject H_0 as there is insufficient evidence to show that P(leading digit = 1) > 0.25

Question by OCR ©

Mean of Normal Distribution $\bar{X} \sim N(\mu, \sigma^2/n)$



This distribution is used with the following scenarios:

- The population data is modelled with $X \sim N(\mu, \sigma^2)$
- A sample of n values has a different mean
- The likeness of this new mean in the original distribution is calculated using $\bar{X} \sim N(\mu, \sigma^2/n)$

H_0: no change in mean (this is almost always H_0)
H_1: mean has decreased
$\alpha = 0.01$ (use ½α in two-tailed tests)

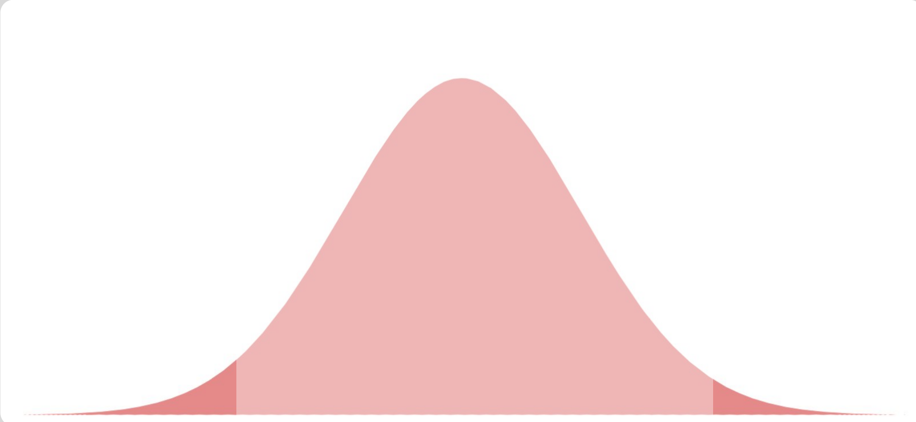
$X \sim N(415, 3.4^2)$
$n = 20$
$\bar{x} = 413$
$\bar{X} \sim N(415, (3.4^2) / 20)$

$P(\bar{X} \leq 413) = 0.004260\ldots \approx 0.0043$
$0.0043 < \alpha$ (if calculated $p < \alpha$ the event is <u>unlikely</u>)

Therefore, reject H_0 and accept H_1 as there is sufficient evidence to suggest that the mean is less than 415

Question by Maths Genie

Standard Normal Distribution $Z \sim N(0, 1)$



This distribution is used with the following scenarios:

- The population data is modelled with $X \sim N(\mu, \sigma^2)$
- You do not know what the mean or the variation is

Set $P(X = x) = P(Z = z)$ and use $z = (x - \mu) / \sigma$ to solve

$X \sim N(\mu, \sigma^2)$
$P(X > 53.28) = 0.0250$
$P(X < 51.65) = 0.0968$

$Z \sim N(0, 1)$
$P(Z > z_1) = 0.0250 \Rightarrow z_1 = 1.959963\ldots \approx 1.96$
$P(Z < z_2) = 0.0968 \Rightarrow z_2 = -1.3$

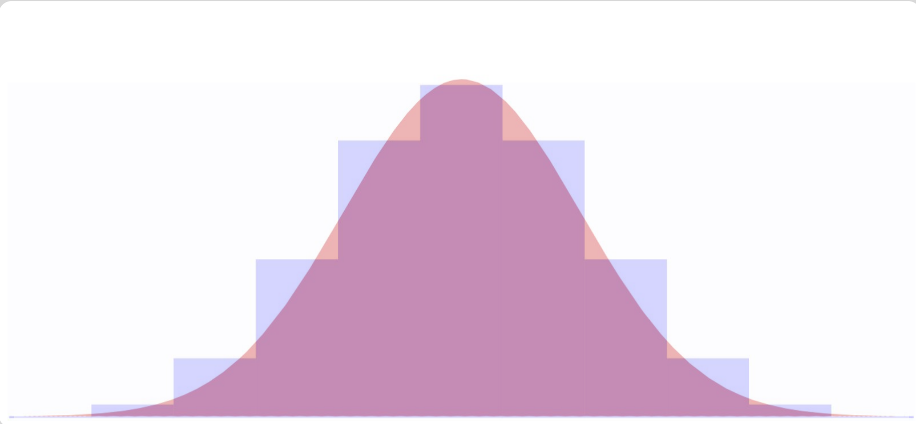
$Z = (X - \mu) / \sigma \Rightarrow \mu = X - \sigma Z$
$\mu = 51.65 - (-1.3)\sigma$
$\mu = 53.28 - (1.96)\sigma$

$51.65 + (1.3)\sigma = 53.28 - (1.96)\sigma$
$(1.3 + 1.96)\sigma = 53.28 - 51.65$
$3.26\sigma = 1.63$
$\sigma = 0.5$

$\mu = 51.65 - (-1.3)\sigma = 51.65 - (-1.3)(0.5)$
$\mu = 52.3$

Question by MadAsMaths

Normal Approximation to Binomial $X \sim N(np, npq)$



This distribution is used if either of the following apply:

- For $X \sim B(n, p)$ the value of $p \approx 0.5$ and n is large
- $np > 5$ and $npq = np(1 - p) > 5$

If the data are not a good fit, the curve will be skewed

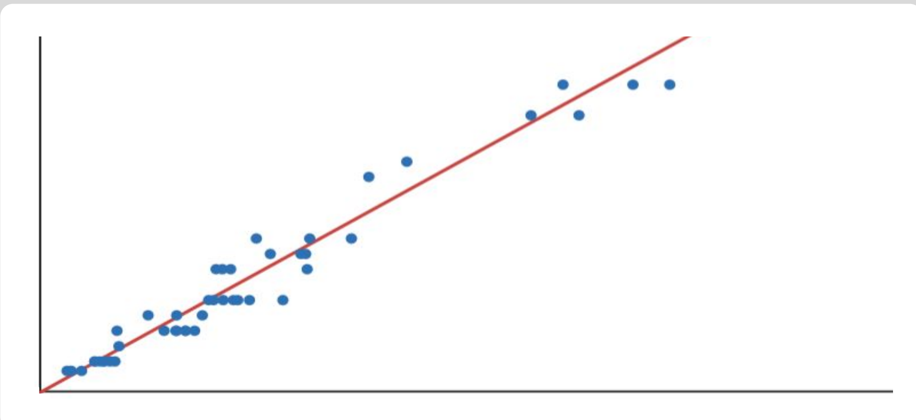
H_0: P(leading digit = 1) = 0.3 (clearly define your P value)
H_1: P(leading digit = 1) > 0.3
$\alpha = 0.05$ (use ½α in two-tailed tests)

$Y \sim B(25, 0.3) \approx X \sim N(7.5, 5.25)$
$25(0.3) = 7.5 > 5$ and $25(0.7)(0.3) = 5.25 > 5$
$P(Y \geq 8) \approx P(X \geq 7.5)$ apply continuity correction
$0.5 > \alpha$ (if calculated $p < \alpha$ the event is unlikely)

Therefore, do not reject H_0 as there is insufficient evidence to show that P(leading digit = 1) $\neq 0.25$

Question by OCR ©

Pearson Product Moment Correlation Coefficient



This is used for plotted data to see how correlated data is:

- If r = 0 there is no correlation
- If r > 0 there is a positive correlation
- If r < 0 there is a negative correlation

No data will likely have a perfect or no correlation

Product Moment Coefficient Level				Sample Size, n
0.10	0.05	0.025	0.01	
…				
0.4716	0.5822	0.6664	0.7498	9
0.4428	0.5494	0.6319	0.7155	10
…				

$r = -0.69$ (this is the correlation factor)
$n = 10$ (this is the sample size)
$H_0: \rho = 0$ (H_0 is always $\rho = 0$)
$H_1: \rho < 0$
$\alpha = 0.05$ (use ½α in two-tailed tests)

Critical Value = 0.5494 and $|r| > 0.5494$

Therefore, reject H_0 as there is sufficient evidence to show that there is a negative correlation

Question by B28 Maths Tutor