Quantitative variable: record amounts and quantities, e.g., number of students in a class, distance between you and your hometown.

Qualitative/Categorical variable: define groups in your data, e.g., students' major in a college, fish species in a lake.

Bar and pie charts can describe categorical data; Histograms and Box plots can depict quantitative data.

The shape of histograms: Skewed to left/right, symmetric, bimodal, multimodal and uniform.

center

Mean: the average value of the data set.

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Median: the middle of the data set when the values of the data set are ranked from smallest to largest.

n is odd: the middle value in the data set

n is even: average of the two middle values in the data set

Mode: the number that appears most often in the data set.

The mean is more sensitive to outliers

The median is less sensitive to outliers

spread

After arranging the data in increasing order, the interquartile range (IQR) is the difference between the 25th percentile (Q_1) and 75th percentile (Q_3) of the data:

$$IQR = Q_3 - Q_1$$

Example: 1, 2, 3, 4, 5, 6, 7.
Median = 4, $Q_1 = 2.5$, $Q_3 = 5.5$, IQR = 3.

Example: 1, 2, 3, 4, 5, 6, 7, 50
Median = 4.5, $Q_1 = 2.5$, $Q_3 = 6.5$, IQR = 4.

The range of the data is defined as the difference between the maximum and minimum values:

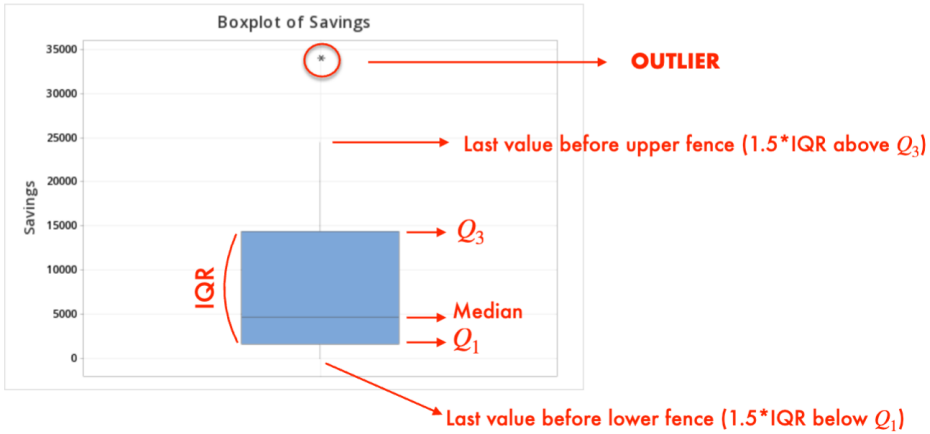
$$range = max - min$$

The variance is the average of the squared differences from the Mean.

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

The standard deviation is a measure of how spread out numbers are.

$$s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$



The z-score:

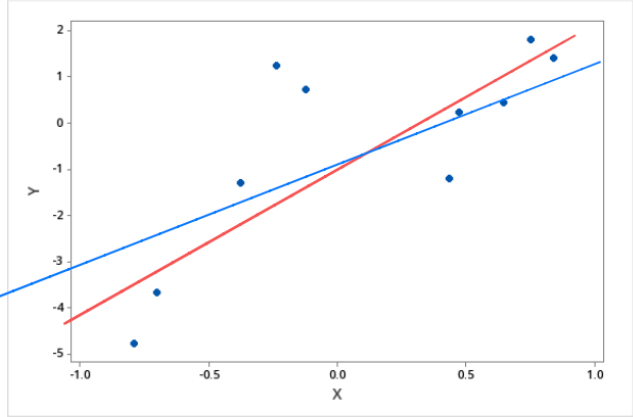
$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Correlation can be used to measure the linear association between two quantitative variables.

Correlation does not imply causation.

Lurking variable ---- We call a hidden variable, behind a relationship and simultaneously affecting the other two variables

We have learned that linear regression can be used to predict the response variable. To perform the linear regression in an appropriate way, we need check the conditions: (1) Quantitative variable; (2) Straight enough; (3) No outliers.



If we focus on a line which has the equation:

$$y = b_0 + b_1 x,$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

$$b_1 = \frac{r s_y}{s_x}$$

when the value for the explanatory variable is x_0, response variable is y_0

Here, we use $\hat{y}$ to denote the predicted response variable, where b_0 is the intercept and b_1 is the slope

For each point $i = 1, \ldots, n$, the vertical distance $y_i - \hat{y}_i$ is called the prediction error or residual

The (sample) correlation between x and y:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

If $r > 0$, then x and y are positively associated
If $r < 0$, then x and y are negatively associated

The (least squares) regression line is the line that minimizes the sum of the squares of the residuals

The coefficient of determination is called R^2 (pronounced "R squared"), which is the fraction of the variation in y explained by the regression

$$R^2 = r^2$$

If $x = \bar{x} + s_x$, meaning the explanatory variable is one standard deviation above the mean, the prediction for the response variable is

We predict y to be r standard deviations above the mean.

The value of R^2 is not usually helpful in determining whether regression is appropriate. It only says how much variation is explained by the regression.

Also know residual plot; extrapolation, curvature, heteroskedasticity, outliers, high leverage points, influential points

Definition of complement event

The **complement** of an event A, denoted A^c, is the event that A does not occur.

Complement rule

We have $P(D^c) = 1 - P(D)$ for all events D.

Definition of disjoint events

Two events are called **disjoint** or **mutually exclusive** if they can not both occur.

Addition rule

If A and B are disjoint, then $P(A \text{ or } B) = P(A) + P(B)$.

In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

Definition of Conditional Probability

Suppose A and B are events and $P(A) > 0$. We define the **conditional probability of B given A** to be

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

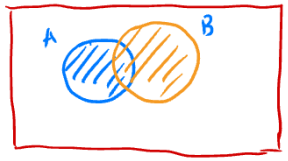
Definition of Independent

Two events A and B are called **independent** if the occurrence of one does not affect the probability that the other occurs. In this case, we have $P(B) = P(B|A)$.

The following three are equivalent:

(1) $P(B) = P(B|A)$ (2) $P(A) = P(A|B)$ (3) $P(A \cap B) = P(A)P(B)$

Venn Diagram



Bayes' Rule

We can reverse the conditional probability by applying the Bayes' rule:

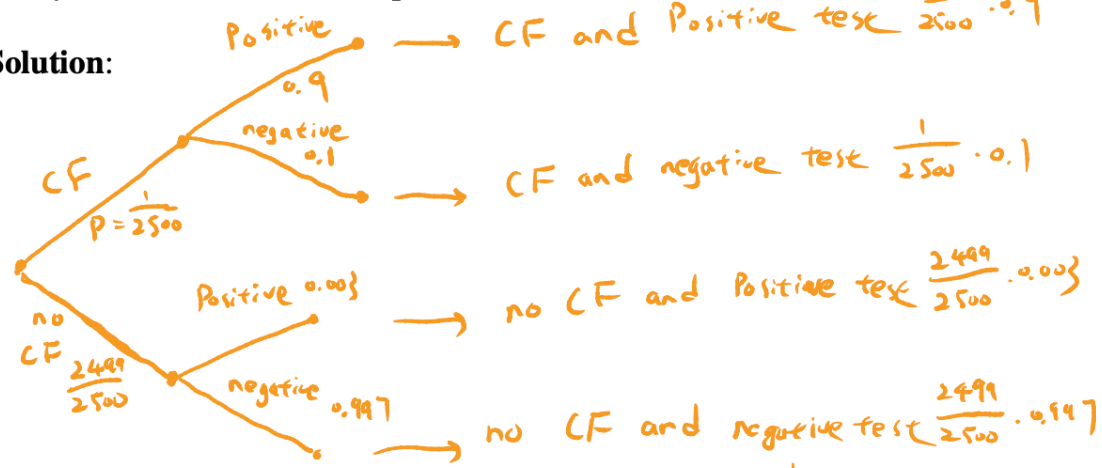
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)}.$$

Simpsons paradox, which is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.

Tree Diagram

Example: A standard test for cystic fibrosis in newborns gives a positive result 90 percent of the time when the baby has cystic fibrosis but gives a false positive 0.3 percent of the time when the baby is healthy. About one baby out of 2500 has cystic fibrosis. What is the probability that the baby has cystic fibrosis if the test is positive?

Solution:



$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)} = \frac{\frac{1}{2500} \cdot 0.9}{P(A \text{ and } B) + P(A \text{ and } B^c)}$$

Definition of random variable

A **random variable** is a quantity whose value is determined by the outcome of a random event or experiment.

	Discrete	Continuous
$P(X = c)$	$P(c)$	0
$P(a \le X \le b)$	$\sum_{i:a \le x_i \le b} P(x_i)$	$\int_a^b f(x)\,dx$
$\mu = E[X]$	$\sum_i x_i P(x_i)$	$\int_{-\infty}^{\infty} x f(x)\,dx$
$\mathrm{Var}(X) = E[(X-\mu)^2]$	$\sum_i (x_i-\mu)^2 P(x_i)$	$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$
alternative way: $\mathrm{Var}(X) = E[X^2] - \mu^2$	$\sum_i x_i^2 P(x_i) - \mu^2$	$\left(\int_{-\infty}^{\infty} x^2 f(x)\,dx\right) - \mu^2$
$\mathrm{SD}(X)$	$\sqrt{\mathrm{Var}(X)}$	$\sqrt{\mathrm{Var}(X)}$

The density f must satisfy:
- $f(x) \ge 0$ for all x.
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Suppose X and Y are random variables and c is a real number.

Properties of Expectation:

- $E[X + c] = E[X] + c$.
- $E[cX] = cE[X]$.
- $E[X + Y] = E[X] + E[Y]$.
- $E[X - Y] = E[X] - E[Y]$

Properties of Variance:

- $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$, so $\mathrm{SD}(cX) = |c|\mathrm{SD}(X)$.
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$
  where $\mathrm{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent.
- $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent.

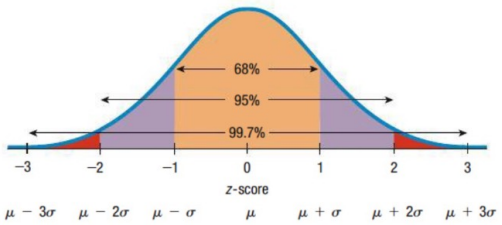
Note: Independent X and Y implies that $\mathrm{Cov}(X, Y) = 0$; however the converse might not be true.

	Description	Support	PDF / density	Mean	Variance
Bernoulli(p)	success (1) with probability p or failure (0) with probability $1-p$	$\{0,1\}$	$P(X=1)=p$ and $P(X=0)=1-p$	p	$p(1-p)$
Binomial(n,p)	# of successes in n independent Bernoulli(p) trials	$k \in \{0,1,\dots,n\}$	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	# of independent Bernoulli(p) trials until the first success	$k \in \{1,2,3,\dots\}$	$P(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	# events occur during a time/space interval (λ is the average # occurrences in the given time/space interval)	$k \in \{0,1,2,\dots\}$	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ
Uniform(a,b)	all intervals of the same length within [a,b] are equally probable	$[a,b]$	$f(x)=\begin{cases}\frac{1}{b-a} & \text{if } a\le x\le b\\ 0 & \text{otherwise}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(λ)	time until the first success (λ is the average # occurrences in the given time/space interval)	$[0,\infty)$	$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{if } x\ge 0\\ 0 & \text{otherwise}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal(μ,σ^2)		$\mathbb{R}$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

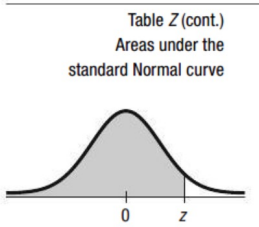
If X is exponential with parameter $\lambda > 0$, then X is a memoryless random variable, that is

$$P(X > x+a \mid X > a) = P(X > x), \quad \text{for } a, x \ge 0.$$

If $X \sim N(\mu,\sigma)$, then $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$.



Know how to read Table Z



	Second decimal place in z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

e.g. $P(Z < 0.56) = 0.7123$

μ and p are **parameters** (fixed numbers that depend on the population).

$\overline{X}$ and $\hat{p}$ are **statistics** (random variables that depend on the sample).

The distributions of $\overline{X}$ and $\hat{p}$ are called **sampling distributions**.

We can make probability statements about statistics, not parameters.

Goal: We want to understand the shape, center, and spread of these distributions, which will help us to understand how accurately $\bar{X}$ and $\hat{p}$ estimate μ and p respectively.

Definition of Law of Large Numbers

As the sample size n tends to infinity, the sample mean (or sample proportion) approaches the population mean (or population proportion). That is, we have

$$\lim_{n\to\infty} \overline{X} = \mu, \qquad \lim_{n\to\infty} \hat{p} = p.$$

Central Limit Theorem

Regardless of the population distribution of $X_1, \ldots, X_n$ coming from, if n is large, the distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

For most distributions (at most moderately skewed), the distribution of $\overline{X}$ is approximately normal if $n \geq 30$.

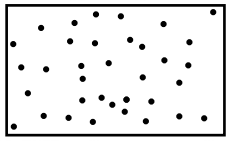
the binomial distribution is approximately $N(np, \sqrt{np(1-p)})$, when $np \geq 10$ and $n(1-p) \geq 10$.

The Central Limit Theorem applies only to averages and sums, not to individual observations.

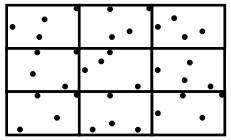
$$\frac{X_1 + \cdots + X_n}{n} \sim_{\text{approx.}} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad X_1 + \cdots + X_n \sim_{\text{approx.}} N(n\mu, \sigma\sqrt{n})$$

Survey/Sampling methods

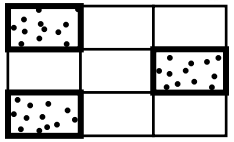
- **Simple random sample**: Choose n people, every sample of size n is equally likely to be chosen.



- **Stratified random sample**: Divide the population into groups called strata, then do simple random sampling in each stratum. (Example: sample 500 men and 500 women rather than any 1000 people.) This can reduce variability.



- **Cluster sample**: Divide the population into clusters. Select a few clusters at random and sample only from those selected. (Example: each dorm is representative of the distribution of majors in the university, then each dorm would be a cluster) This can reduce costs.



- **Voluntary response sample**: Many people are invited to respond, and all who respond are counted. (Example: surveys done through the internet or radio talk shows.) These surveys suffer from voluntary response bias, and have no scientific value.

Poisson approximation

When n is large and p is small, the Binomial(n, p) distribution can be approximated by the Poisson distribution with $\lambda = np$.

Rule of thumb: the Poisson model is a reasonably good approximation of the Binomial when $n \geq 20$ with $p \leq 0.05$ or $n \geq 100$ with $p \leq 0.10$

Confidence Interval

- CI is an interval of plausible values that contains the true parameter value with a specific probability, we call this probability to be Confidence Level.
- We are 95 percent confident that our confidence interval contains p, in the sense that if we took many samples of size 1000 and repeated this process, 95 percent of the intervals would contain p.

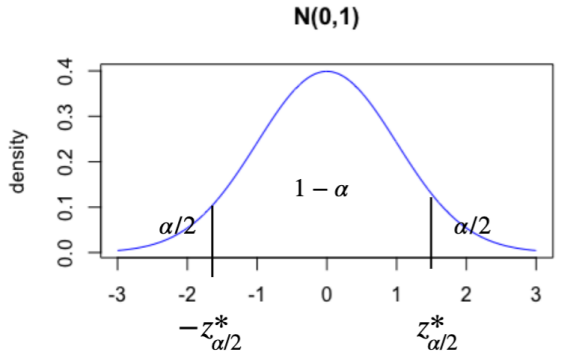
<u>One-proportion z-interval</u>

Suppose n is the sample size p is the population proportion, $\hat{p}$ is the sample proportion and $0 < \alpha < 1$ is the significance level.

Then, the $(1-\alpha)$ confidence interval is

$$\left[\hat{p} - z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

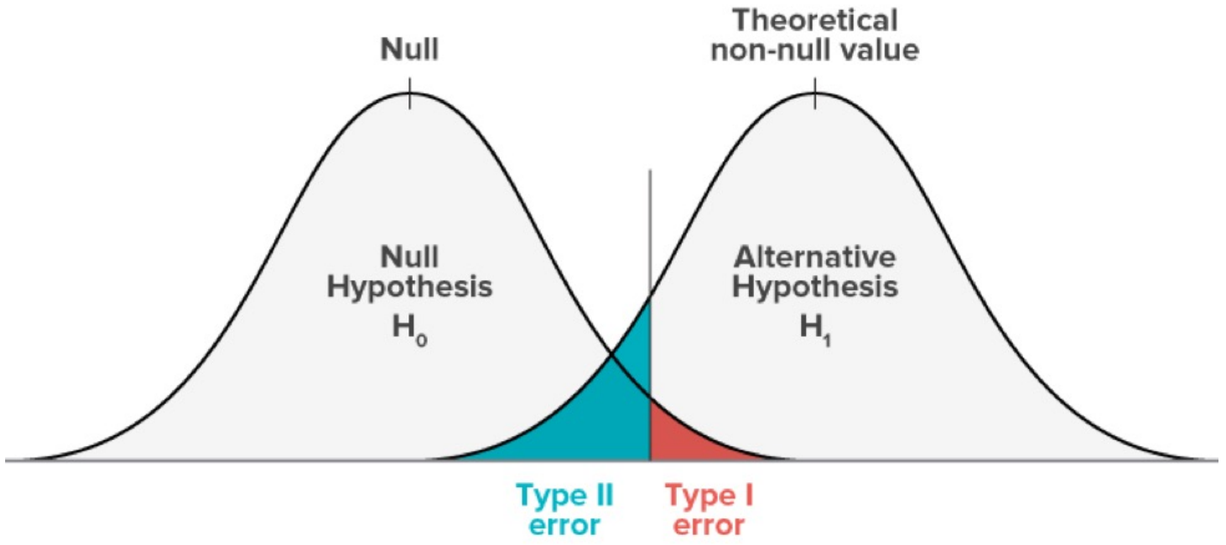
The margin of error is $z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.



Type I Error: Reject H_0 when H_0 is true. This happens with probability equal to the significance level α.

Type II Error: Fail to reject H_0 when H_0 is false. We denote by β the probability of a Type II error, which depends on the value of p (the true proportion). The **power** of the test is $1-\beta$, which is the probability of rejecting H_0 when H_0 is indeed false.

Decision	True: H_0 is true	True: H_A is true
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision



$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{Fail to reject } H_0 \mid H_A \text{ is true})$$

Ideally we would like both α, β to be small

Need to know these four
(the other two are here for completeness)

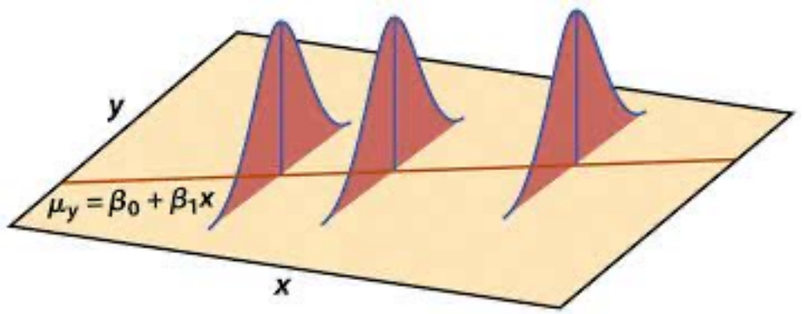
	True Parameter(s)	Sample Statistic(s)	Approximate Distribution (From Central Limit Theorem)		Confidence Interval	Hypothesis Testing Test Statistics
One Proportion	p	$\hat{p}$	$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$		$\hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}}$
not for this class Two Proportions	p_1, p_2	$\hat{p}_1, \hat{p}_2$	$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$		$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_*(1-p_*)}{n_1} + \frac{p_*(1-p_*)}{n_2}}}$ Pooled Proportion Here
One Mean	μ	$\bar{x}$	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim Z = N(0,1)$	But unknown true variance σ $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim T_{n-1}$	$\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$	$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
Two Means (paired)	μ_d	$\bar{x}_d$	$\frac{\bar{X}_d - \mu_d}{\sigma_d/\sqrt{n}} \sim Z$	But unknown true variance σ_d $\frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}} \sim T_{n-1}$	$\bar{x}_d \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$ s_d can hardly be obtained by s_1 and s_2	$t_{n-1} = \frac{\bar{x}_d - 0}{\frac{s_d}{\sqrt{n}}}$
not for this class Two Means (independent, equal variance)	μ_1, μ_2	$\bar{x}_1, \bar{x}_2$	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim Z$ (note that $\sigma = \sigma_1 = \sigma_2$)	But unknown true variance σ $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2}$ (where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$)	$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Two Means (independent, unequal variance)	μ_1, μ_2	$\bar{x}_1, \bar{x}_2$	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z$	But unknown true variance σ_1, σ_2 $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T_{df}$ For this class, we use $df = \min(n_1 - 1, n_2 - 1)$ for simplicity and as a **conservative** approximation (smaller df >>> bigger tail in T_{df} distribution)	$(\bar{x}_1 - \bar{x}_2) \pm t_{\min(n_1-1,n_2-1),\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$t_{\min(n_1-1,n_2-1)} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \geq \min(n_1 - 1, n_2 - 1)$$

Regression Inference

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

We usually assume that ϵ_i has 0 mean and finite variance σ.



Know how to read

Minitab outputs ★

$b_0 = \bar{y} - b_1\bar{x}$

$b_1 = \dfrac{rs_y}{s_x}$

Term	Coef	SE Coef	T-Value	P-Value
Constant	-35.075	1.832	-19.15	0.000
Year	0.0178	0.000918	19.39	0.000

$$s_x = \sqrt{\frac{(x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}}$$

$$s_y = \sqrt{\frac{(y_1-\bar{y})^2 + \cdots + (y_n-\bar{y})^2}{n-1}}$$

$$\mathrm{SE}(b_1) = \frac{s}{s_x\sqrt{n-1}}$$

where $\hat{y}_i = b_0 + b_1 x_i$ $\quad s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

$$T = \frac{b_1}{\mathrm{SE}(b_1)}$$

with $n-2$ degrees of freedom

Confidence interval for slope $\quad (b_1 - t^*_{n-2,\frac{\alpha}{2}}\mathrm{SE}(b_1),\, b_1 + t^*_{n-2,\frac{\alpha}{2}}\mathrm{SE}(b_1))$

Hypothesis testing for slope $\quad$ Test $H_0 : \beta_1 = 0, H_A : \beta_1 \neq 0$ $\quad$ Test statistic $\mathrm{T} = \dfrac{b_1}{SE(b_1)}$

★

Fit	95% CI	95% PI
69.8751	(69.6838, 70.0665)	(65.0844, 74.6659)

A 95% **confidence interval** for the mean response $\mu_y = \beta_0 + \beta_1 x^*$.

not for this class $\quad SE(\mu_y) = \sqrt{\frac{s^2}{n} + (x^* - \bar{x})^2 \cdot SE^2(b_1)}$ $\quad$ CI is $\left[\mu_y - t^*_{n-2,\frac{\alpha}{2}}SE(\mu_y), \mu_y + t^*_{n-2,\frac{\alpha}{2}}SE(\mu_y)\right]$

A 95% **prediction interval** for an individual response $y^* = \beta_0 + \beta_1 x^* + \varepsilon^*$. $\quad \hat{y} = \beta_0 + \beta_1 x^*$

not for this class $\quad SE(\hat{y}) = \sqrt{s^2 + \frac{s^2}{n} + (x^* - \bar{x})^2 \cdot SE^2(b_1)}$ $\quad$ PI is $\left[\hat{y} - t^*_{n-2,\frac{\alpha}{2}}SE(\hat{y}), \hat{y} + t^*_{n-2,\frac{\alpha}{2}}SE(\hat{y})\right]$

Chi-square Tests

If $Z_1, \ldots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

has the chi-squared distribution with k degrees of freedom, which usually denoted as $Q \sim \chi^2(k)$.

Three different kinds of chi-square tests:

- Chi-square test for goodness-of-fit.
- Chi-square test for homogeneity.
- Chi-square test for independence.

The chi-square test for goodness-of-fit is used when we have one categorical variable. The other two tests are used when we want to determine whether two categorical variables are associated.

Chi-square test for goodness-of-fit: to test whether categorical data are generated from a specified probability distribution.

Example:

Type	Observed	Expected
round, yellow	315	(556)(9/16) = 312.75
round, green	108	(556)(3/16) = 104.25
wrinkled, yellow	101	(556)(3/16) = 104.25
wrinkled, green	32	(556)(1/16) = 34.75

H_0: In this experiment, 9/16 of peas should be round and yellow, 3/16 should be round and green, 3/16 should be wrinkled and yellow, and 1/16 should be wrinkled and green.

H_A: The four types of peas arise with some other probabilities.

$$\chi^2 = \sum_{\text{categories}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \cdots \approx 0.47.$$

$$df = (\text{categories} - 1)$$

p-value is $P\left(\chi^2_{df} \geq 0.47\right) \approx 0.93$

- **Chi-square test for homogeneity**: to compare the distributions of two or more groups for the same categorical variable.
- **Chi-square test for independence**: to test whether two categorical variables are independent.

The mechanics are exactly the same, regardless of whether we are testing homogeneity or independence.

Example:

Observed Counts

	Boys	Girls	Total
Grades	117	130	247
Popular	50	91	141
Sports	60	30	90
Total	227	251	478

The expected count in a cell is $\frac{\text{Row Total} \times \text{Column Total}}{\text{Table Total}}$.

$247 \times 227 / 478 = 117.3$

Expected Counts

	Boys	Girls
Grades	117.3	129.7
Popular	67.0	74.0
Sports	42.7	47.3

H_0: boys and girls are the same in how they prioritize grades, popularity, and sports.

H_A: boys and girls prioritize grades, popularity, and sports differently.

$$\chi^2 = \sum_{\text{rows}} \sum_{\text{columns}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \cdots \approx 21.46.$$

$$df = (\text{rows} - 1)(\text{columns} - 1)$$

p-value is $P\left(\chi^2_{df} \geq 21.46\right) \approx 0.000022$