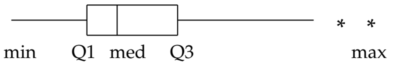
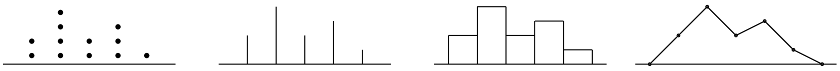
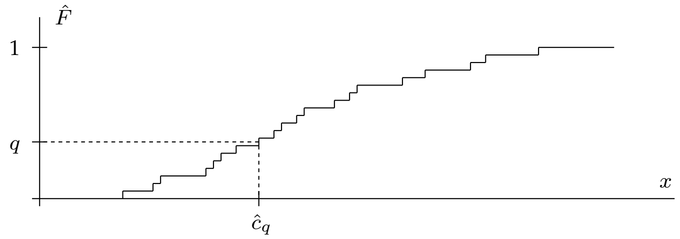
Experimental Design & Data Analysis: Summary notes

STATISTICS	
Types of variable	*properties*
categorical	category
ordinal	category + order
numerical	category + order + scale; [counting = discrete, measurement = continuous]
Descriptive statistics	for $\{x_1, x_2, \dots, x_n\}$; order statistics $(x_{(1)} \leqslant x_{(2)} \leqslant \dots \leqslant x_{(n)})$.
sample mean, $\bar{x}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \quad \approx \quad \frac{1}{n}\sum_{j=1}^k f_j u_j$.
sample median, $\hat{m}$, $\hat{c}_{0.5}$	the middle observation, $x_{(\frac{1}{2}(n+1))}$
sample P-trimmed mean	trim off $\lceil \frac{1}{2}nP \rceil$ observations at each end, and average the rest.
sample mid-range	$\frac{1}{2}(x_{(1)} + x_{(n)})$
sample mode, $\hat{\text{M}}$	the most frequent observation, or the midpoint of the most frequent class.
sample quantile, $\hat{c}_q$	$\hat{c}_q = x_{(k)}$, where $k = (n+1)q$.
sample quartiles	Q1 $= \hat{c}_{0.25}$, Q3 $= \hat{c}_{0.75}$ (Q2 $= \hat{m} = \hat{c}_{0.5}$).
five-number summary	(min, Q1, med, Q3, max)
boxplot	 'outliers' outside $(\text{Q1} - 1.5\,\text{IQR}, \text{Q3} + 1.5\,\text{IQR})$
sample variance, s^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
form for computation	$= \frac{1}{n-1}\big(\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2\big) \quad \approx \quad \frac{1}{n-1}\big(\sum_{j=1}^k f_j u_j^2 - \frac{1}{n}(\sum_{j=1}^k f_j u_j)^2\big)$
sample standard deviation, s	$\sqrt{s^2}$
sample interquartile range, IQR	IQR $=$ Q3 $-$ Q1, $\hat{\tau} = \hat{c}_{0.75} - \hat{c}_{0.25}$ (a number, not an interval)
sample range	$x_{(n)} - x_{(1)}$
sample skewness	$\hat{\lambda}_3 = \hat{\nu}_3 / s^3$, where $\hat{\nu}_3 = \frac{1}{n-2}\sum(x_i - \bar{x})^3$
sample kurtosis	$\hat{\lambda}_4 = \hat{\nu}_4 / s^4 - 3$, where $\hat{\nu}_4 = \frac{1}{n-3}\sum(x_i - \bar{x})^4$
frequency distributions	dotplot bar graph histogram frequency polygon
	
sample pmf, $\hat{p}(x)$	$\hat{p}(x) = \frac{1}{n}\,\text{freq}(X = x)$
sample pdf, $\hat{f}(x)$	$\hat{f}(x) = \frac{1}{b-a}\,\text{freq}(a < X < b)$ for cell $a < x < b$ [histogram]
sample cdf, $\hat{F}(x)$	$\hat{F}(x) = \frac{1}{n}\,\text{freq}(X \leqslant x)$; $\hat{F}(x) = \frac{k}{n}$, $(x_{(k)} \leqslant x < x_{(k+1)})$
	
sample quantiles (inverse cdf)	$\hat{F}(\hat{c}_q) \approx q$; $\hat{c}_q \approx \hat{F}^{-1}(q)$.
sample covariance, s_{xy}	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
sample correlation, $r = r_{xy}$	$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\ \Sigma(y-\bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^n x_{si} y_{si}$.
risk (incidence proportion) R	$\hat{R} = \frac{\text{number developing disease } D \text{ during time period } \Delta t}{\text{number of individuals followed for the time period}}$
incidence rate, α	$\hat{\alpha} = \frac{\text{number of individuals developing disease } D \text{ in a time interval}}{\text{total time for which individuals were followed}}$
prevalence proportion, π	$\hat{\pi} = \frac{\text{number of individuals with characteristic } D \text{ at time } t}{\text{total number of individuals}}$

STATISTICS

Data sources. Types of studies:

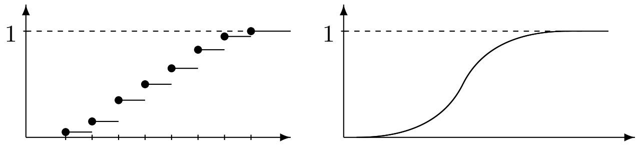
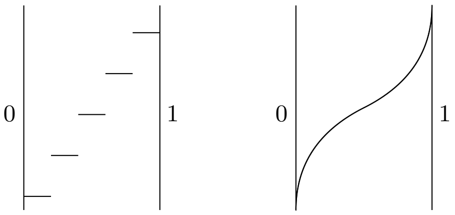
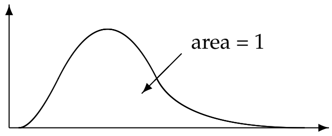
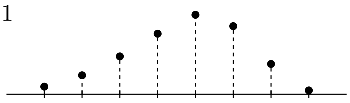
experimental studies	observational studies
clinical trials	cohort (follow-up, prospective)
field trials	case-control (retrospective)
community intervention	cross sectional (survey)
imposed intervention (randomisation)	no intervention
inferred causation	no inferred causation

statistical experiments:	treatments applied to experimental units and their effect on the response variable is observed
desirable qualities of an experiment:	(1) validity (unbiasedness); (2) precision (efficiency).
validity control group	no treatment; placebo = simulated (non)treatment
randomisation	each unit has an equal probability of being assigned each treatment
precision blocking (stratification)	a block is a group of similar experimental units; block ≈ sub-experiment: randomise within blocks
replication	more observations increases precision
balance	balance is preferable: i.e. equal numbers with each treatment
confounding variable	an explanatory variable whose effect distorts the effect of another.
lurking variable	an unobserved variable that could be a confounding variable

PROBABILITY, Pr	(a set function defined on an event space)
random experiment	a procedure leading to an observable outcome
event space, Ω	set of possible outcomes
event, A	subset of event space
properties of probability function (1)	$0 \leqslant \Pr(A) \leqslant 1$ for all events A
(2)	$\Pr(\emptyset) = 0, \Pr(\Omega) = 1$
(3)	$\Pr(A') = 1 - \Pr(A)$ (A' denotes the complement of A).
(4)	$A \subseteq B \;\Rightarrow\; \Pr(A) \leqslant \Pr(B)$
(5)	$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ [addition theorem]
assigning values to Pr	*symmetry; long-term relative frequency; subjective; model*
odds, $\mathcal{O}$	$\mathcal{O}(A) = \frac{\Pr(A)}{\Pr(A')}$; odds = $\frac{p}{1-p}$
probability table for A and B	$\begin{array}{c\|cc\|c} & B & B' & \\ \hline A & \Pr(A{\cap}B) & \Pr(A{\cap}B') & \Pr(A) \\ A' & \Pr(A'{\cap}B) & \Pr(A'{\cap}B') & \Pr(A') \\ \hline & \Pr(B) & \Pr(B') & 1 \end{array}$ $\boxed{\begin{array}{cc} \alpha & \beta \\ \gamma & \delta \end{array}}$
conditional probability	$\Pr(A \mid H) = \frac{\Pr(A{\cap}H)}{\Pr(H)}, \quad \Pr(H) \neq 0$
conditional odds	$\mathcal{O}(A \mid H) = \frac{\Pr(A \mid H)}{\Pr(A' \mid H)}$.
multiplication rule	$\Pr(A{\cap}B) = \Pr(A)\Pr(B \mid A) = \Pr(B)\Pr(A \mid B)$
relationship between A and B	$\Pr(A \mid B) \gtrless \Pr(A) \gtrless \Pr(A \mid B')$ (positive relationship negative relationship)
law of total probability	$\Pr(H) = \sum_{i=1}^{m} \Pr(A_i)\Pr(H \mid A_i)$ for $\{A_i\}$ a partition of Ω.
Bayes' theorem	$\Pr(A_k \mid H) = \frac{\Pr(A_k)\Pr(H \mid A_k)}{\sum_{i=1}^{m}\Pr(A_i)\Pr(H \mid A_i)}$ for $\{A_i\}$ a partition of Ω. mutually exclusive and exhaustive "causes" $A_1, A_2, \ldots, A_k$ of "result" H *e.g. exposure → disease; disease → test result*
relative risk (risk ratio), RR	$\text{RR} = \frac{\Pr(D \mid E)}{\Pr(D \mid E')}$ for disease D with exposure E; $\text{RR} = \frac{\alpha(\gamma+\delta)}{\gamma(\alpha+\beta)}$
odds ratio, OR	$\text{OR} = \frac{\mathcal{O}(D \mid E)}{\mathcal{O}(D \mid E')}$ for disease D with exposure E; $\text{OR} = \frac{\alpha\delta}{\beta\gamma}$

<u>*Diagnostic testing*</u>	D = individual has disease, P = individual tests positive
sensitivity	sn = $\Pr(P \mid D)$
specificity	sp = $\Pr(P' \mid D')$
positive predictive value	ppv = $\Pr(D \mid P)$
negative predictive value	npv = $\Pr(D' \mid P')$
errors	false positive = $D'{\cap}P$; false negative = $D{\cap}P'$
prevalence, prior probability	$\Pr(D)$

Independent events	$\Pr(A\cap B)=\Pr(A)\Pr(B)\neq 0$ (e.g. H_1, H_2)
cf. mutually exclusive events	$A\cap B=\emptyset,\ \Pr(A\cap B)=0$ (e.g. H_1, T_1)
independence of n events	$\Pr(A_{j_1}\cap A_{j_2}\cap\cdots\cap A_{j_m})=\Pr(A_{j_1})\Pr(A_{j_2})\cdots\Pr(A_{j_m})$
if $A_1, A_2, \ldots, A_n$ independent, then:	$\Pr(A_1\cap A_2\cap\cdots\cap A_n)=\Pr(A_1)\Pr(A_2)\cdots\Pr(A_n)$ $\Pr(A_1\cup A_2\cup\cdots\cup A_n)=1-\Pr(A_1'\cap\cdots\cap A_n')=\Pr(A_1')\cdots\Pr(A_n')$, i.e. Pr("*at least one*") = 1 – Pr("*none*").

Random variable, $X\colon\Omega\to\mathbb{R}$	*Maths defn: real-valued function defined on* Ω, $X(\omega), \omega\in\Omega$. a numerical outcome of a random procedure.
sample space, S	the set of possible values of X, i.e. the range of the function $X\colon\Omega\to S\subseteq\mathbb{R}$
cumulative distribution function, cdf	$F(x)=\Pr(X\leqslant x)$
nsc for F to be a cdf (1)	F non-decreasing
(2)	$F(-\infty)=0,\ F(\infty)=1$
(3)	F right-continuous, i.e. $F(x+0)=F(x)$.
sketch cdf	
probability from cdf	$\Pr(a<X\leqslant b)=F(b)-F(a)$
sketch inverse cdf, F^{-1}	
q-quantile, c_q $(0<q<1)$	$c_q=F_X^{-1}(q)$
continuous random variables	$\Pr(X=x)=0$
probability density function, pdf	$f(x)=\frac{d}{dx}\big(F(x)\big)$; $\Pr(X\approx x)\approx f(x)\delta x$
nsc for f to be a pdf (1)	$f(x)\geqslant 0$
(2)	$\int_{-\infty}^{\infty}f(x)dx=1$
probability from pdf	$\Pr(a<X\leqslant b)=\int_a^b f(x)dx \ \Rightarrow\ F(x)=\int_{-\infty}^{x}f(t)dt$
sketch pdf	
discrete random variables	
probability mass function, pmf	$p(x)=\Pr(X=x)$
nsc for p to be a pmf (1)	$p(x)\geqslant 0$
(2)	$\sum p(x)=1$
sketch pmf	
relation of pmf to cdf	$p(x)=F(x+0)-F(x-0)=$ jump in F at x

Expectation, E	
expectation of $\psi(X)$	$\mathrm{E}(\psi(X))=\int\psi(x)f(x)dx$ or $\sum\psi(x)p(x)$
mean of X, μ, $\mathrm{E}(X)$	$\int xf(x)dx$ or $\sum xp(x)$
$\mathrm{E}(a+bX)$, $\mathrm{E}(X+Y)$	$a+b\,\mathrm{E}(X)$, $\mathrm{E}(X)+\mathrm{E}(Y)$
median of X, m	0.5-quantile, $c_{0.5}=F^{-1}(0.5)$
mode of X, M	$f(\mathrm{M})\geqslant f(x)$ for all x or $p(\mathrm{M})\geqslant p(x)$ for all x
variance of X, $\mathrm{var}(X)$, σ^2	$\mathrm{E}\big((X-\mu)^2\big)=\mathrm{E}(X^2)-\mathrm{E}(X)^2$
standard deviation, $\mathrm{sd}(X)$, σ	$\mathrm{sd}(X)=\sqrt{\mathrm{var}(X)}$
$\mathrm{var}(a+bX)$, $\mathrm{sd}(a+bX)$	$b^2\mathrm{var}(X)$, $\lvert b\rvert\,\mathrm{sd}(X)$
$\mathrm{var}(X+Y)$ *(X and Y independent)*	$\mathrm{var}(X)+\mathrm{var}(Y)$
approxns for mean & sd of $g(X)$	$\mathrm{E}[g(X)]\approx g(\mu)$, $\mathrm{sd}[g(X)]\approx\lvert g'(\mu)\rvert\,\mathrm{sd}(X)$, provided $\mathrm{sd}(X)$ small.
covariance of X and Y, $\mathrm{cov}(X,Y)$	$\sigma_{XY}=\mathrm{E}\big((X-\mu_X)(Y-\mu_Y)\big)$ (zero if X and Y are independent).
correlation of X and Y, $\rho(X,Y)$	$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$ (zero if X and Y are independent).
$\mathrm{var}(aX+bY)$	$a^2\mathrm{var}(X)+b^2\mathrm{var}(Y)+2ab\,\mathrm{cov}(X,Y)$

Linear combinations of independent rvs	$Y = a_1X_1 + a_2X_2 + \cdots + a_kX_k$, with $\mathrm{E}(X_i) = \mu_i$, $\mathrm{var}(X_i) = \sigma_i^2$	
mean of $a_1X_1{+}a_2X_2{+}\cdots{+}a_kX_k$	$\mathrm{E}(Y) = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k$,	$\mathrm{E}(X_1{-}X_2) = \mu_1{-}\mu_2$;
variance of $a_1X_1{+}a_2X_2{+}\cdots{+}a_kX_k$	$\mathrm{var}(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_k^2\sigma_k^2$,	$\mathrm{var}(X_1{-}X_2) = \sigma_1^2{+}\sigma_2^2$;
if $X_1, X_2, \ldots, X_k$ normally distributed	then $Y = a_1X_1{+}a_2X_2{+}\cdots{+}a_kX_k$ is normally distributed.	
combining indept unbiased estimators	$T_1, T_2, \ldots, T_k$ independent, with $\mathrm{E}(T_i) = \theta$ and $\mathrm{var}(T_i) = \sigma_i^2$.	
optimal $T = a_1T_1{+}\cdots{+}a_kT_k$	$a_i = \frac{c}{\sigma_i^2}$, where $c = 1/(\frac{1}{\sigma_1^2} + \cdots + \frac{1}{\sigma_k^2}) \;\Rightarrow\; \mathrm{E}(T) = \theta,\ \mathrm{var}(T) = c$.	
Random sampling: iidrvs	independent identically distributed random variables	
random sample on X	$X_1, X_2, \ldots, X_n$ iidrvs $\overset{\mathrm{d}}{=} X$	
statistic, T	$T = \psi(X_1, X_2, \ldots, X_n)$	
distribution of frequencies	freq$(A) \overset{\mathrm{d}}{=} \mathrm{Bi}(n, \Pr(A))$	
sample mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$	$\mathrm{E}(\bar{X}) = \mu,\ \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$
sample variance, S^2	$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$	$\mathrm{E}(S^2) = \sigma^2$.
law of large numbers	If $\mu = \mathrm{E}(X) < \infty$ then $\bar{X} \xrightarrow{\mathrm{p}} \mu$ as $n \to \infty$	
central limit theorem	If also $\sigma^2 = \mathrm{var}(X) < \infty$, then $\bar{X} \overset{\mathrm{d}}{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n})$	
Statistical Inference estimator of θ	T is a statistic chosen so that it will be close to θ	
estimate of θ	t is a realisation of an estimator T	
unbiasedness (for θ)	$\mathrm{E}(T) = \theta$	
Confidence interval		"basic confidence interval": est ± "2"se
confidence interval for θ based on T	realisation of the random interval $\big(\ell(T), u(T)\big)$, where $\Pr(\ell(T) < \theta < u(T)) = \gamma$; CI for θ: $\big(\ell(t), u(t)\big)$	
Hypothesis testing		"basic test statistic": $\frac{\text{est}-\theta_0}{\text{se}^*}$, cf. "2"
significance level	$\alpha = \Pr(\text{reject } H_0 \mid H_0)$,	$\alpha = \Pr(\text{type I error}) = \Pr(R \mid H_0)$
power	$1{-}\beta = \Pr(\text{reject } H_0 \mid H_1))$	$\beta = \Pr(\text{type II error}) = \Pr(R' \mid H_1)$
power function	$Q(\theta) = \Pr(\text{reject } H_0 \mid \theta)$	
p-value	$\Pr(\text{test statistic is at least as extreme as the value observed} \mid H_0)$; reject H_0 if $\mathbf{p} < \alpha$.	
Inference for normal populations	(variance known)	(variance unknown)
one sample: n on $\mathrm{N}(\mu, \sigma^2)$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\mathrm{d}}{=} \mathrm{N}$	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-1}$
100(1−α)% CI for μ	$\bar{x} \pm c_{1-\frac{1}{2}\alpha}(\mathrm{N})\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm c_{1-\frac{1}{2}\alpha}(\mathrm{t}_{n-1})\frac{s}{\sqrt{n}}$
100(1−α)% PI for X	$\bar{x} \pm c_{1-\frac{1}{2}\alpha}(\mathrm{N})\,\sigma\sqrt{1+\frac{1}{n}}$	$\bar{x} \pm c_{1-\frac{1}{2}\alpha}(\mathrm{t}_{n-1})\,s\sqrt{1+\frac{1}{n}}$
test statistic for $\mu{=}\mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
sample size calculations 100(1−α)% CI = [est ± w];	$n \geqslant \frac{z_{1-\frac{1}{2}\alpha}^2\sigma^2}{w^2}$;	
sig level (μ_0) α; power (μ_1) $1{-}\beta$	$n \geqslant \frac{(z_{1-\frac{1}{2}\alpha}+z_{1-\beta})^2\sigma^2}{(\mu_1-\mu_0)^2}$;	and if $\sigma_1{\neq}\sigma_0$: $n \geqslant \frac{(z_{1-\frac{1}{2}\alpha}\sigma_0 + z_{1-\beta}\sigma_1)^2}{(\mu_1-\mu_0)^2}$
checking Normality: QQ plot	$\{(\Phi^{-1}(\frac{k}{n+1}), x_{(k)}),\ k = 1, 2, \ldots, n\}$;	
if Normal model is correct	points should be close to a straight line with intercept μ and slope σ.	
probability plot for Normality	QQ plot with axes interchanged [and $\Phi^{-1}(q)$ relabelled as q.]	
two samples: n_1 on $\mathrm{N}(\mu_1, \sigma_1^2)$ n_2 on $\mathrm{N}(\mu_2, \sigma_2^2)$	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \overset{\mathrm{d}}{=} \mathrm{N}$;	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{\mathrm{d}}{\approx} \mathrm{t}_k$; where $\min(n_1{-}1, n_2{-}1) \leqslant k \leqslant n_1{+}n_2{-}2$.
100(1−α)% CI for $\mu_1{-}\mu_2$	$\bar{x}_1{-}x_2 \pm c_{1-\frac{1}{2}\alpha}(\mathrm{N})\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\bar{x}_1{-}x_2 \pm c_{1-\frac{1}{2}\alpha}(\mathrm{t}_k)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
test statistic for $\mu_1{-}\mu_2 = 0$	$z = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \overset{\mathrm{d}}{=} \mathrm{N}$;	$t^* = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \overset{\mathrm{d}}{=} \mathrm{t}_k$;

if $\sigma_1^2=\sigma_2^2=\sigma^2$, then	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S^2(\frac{1}{n_1}+\frac{1}{n_2})}} \overset{\mathrm{d}}{=} \mathrm{t}_{n_1+n_2-2}$, where $S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.
100(1−α)% CI for $\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2 \pm c_{1-\frac{1}{2}\alpha}(\mathrm{t}_{n_1+n_2-2})\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}$
test statistic for $\mu_1=\mu_2$	$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$
sample size calculations	
100(1−α)% CI = [est ± w];	$n_1=n_2 \geqslant \frac{2z^2_{1-\frac{1}{2}\alpha}\sigma^2}{w^2}$;
sig level α; power(d) = 1−β	$n_1=n_2 \geqslant \frac{2(z_{1-\frac{1}{2}\alpha}+z_{1-\beta})^2\sigma^2}{d^2}$ $\{Z=\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{2}{n}}},\ \theta=\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{2}{n}}}\}$
Rank test (for location)	replace observations by ranks: $\frac{\bar{W}_1-\bar{W}_2}{\sigma_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{\mathrm{d}}{\approx}$ N, where $\sigma_W^2=\frac{1}{12}(n_1+n_2)(n_1+n_2+1)$.
Inference for proportions	
one sample of n	$\hat{p}=\frac{x}{n}$; $X \overset{\mathrm{d}}{=} \mathrm{Bi}(n,p) \approx \mathrm{N}\big(np,np(1-p)\big)$ $(np{>}5, nq{>}5)$ [CC]
large n	est $=\hat{p}$, se $=\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, se$_0=\sqrt{\frac{p_0(1-p_0)}{n}}$ CI: est ± $z_{1-\frac{1}{2}\alpha}$ se; HT: $z=\frac{\text{est}-p_0}{\text{se}_0}$
small n	MINITAB, Statistic-Parameter diagram [Figure 2]
testing median, $m=m_0$	equivalent to testing $p=\Pr(X<m_0)=0.5$; $\mathrm{H}_0(m=m_0) \Rightarrow \hat{p}=\frac{u}{n}$, where $U \overset{\mathrm{d}}{=} \mathrm{Bi}(n,0.5)$
two samples of n_1 and n_2	$X_i \overset{\mathrm{d}}{=} \mathrm{Bi}(n_i,p_i) \approx \mathrm{N}\big(n_ip_i,n_ip_i(1-p_i)\big)$; $\hat{p}_i=\frac{x_i}{n_i}$.
large n confidence interval	est $=\hat{p}_1-\hat{p}_2$, se $=\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$; CI: est ± $z_{1-\frac{1}{2}\alpha}$ se;
large n test $p_1=p_2$	est $=\hat{p}_1-\hat{p}_2$, se$_0=\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}$, $\hat{p}=\frac{x_1+x_2}{n_1+n_2}$; HT: $z=\frac{\text{est}}{\text{se}_0}$
sample size calculations	use $\sigma_0^2=p_0(1-p_0)$ and $\sigma_1^2=p_1(1-p_1)$ in the Normal results above $(\sigma_0 \neq \sigma_1)$.
Inference for rates	
one sample for person-time t	$\hat{\alpha}=\frac{x}{t}$; $X \overset{\mathrm{d}}{=} \mathrm{Pn}(\alpha t) \approx \mathrm{N}(\alpha t,\alpha t)$ $(\alpha t{>}10)$ [CC]
large t	est $=\hat{\alpha}$, se $=\sqrt{\frac{\hat{\alpha}}{t}}$; se$_0=\sqrt{\frac{\alpha_0}{t}}$; CI: est ± $z_{1-\frac{1}{2}\alpha}$ se; HT: $z=\frac{\text{est}-\alpha_0}{\text{se}_0}$
small t	MINITAB, Statistic-Parameter diagram [Figure 4]
expected number of cases, λ	$\hat{\lambda}=x$; X, number of cases $\overset{\mathrm{d}}{=} \mathrm{Pn}(\lambda) \approx \mathrm{N}(\lambda,\lambda)$ $(\lambda>10)$ [CC]
two samples for t_1 and t_2	$X_i \overset{\mathrm{d}}{=} \mathrm{Pn}(\alpha_it_i) \approx \mathrm{N}\big(\alpha_it_i,\alpha_it_i\big)$; $\hat{\alpha}_i=\frac{x_i}{t_i}$.
large t confidence interval	est $=\hat{\alpha}_1-\hat{\alpha}_2$, se $=\sqrt{\frac{\hat{\alpha}_1}{t_1}+\frac{\hat{\alpha}_2}{t_2}}$; CI: est ± $z_{1-\frac{1}{2}\alpha}$ se;
large t test $\alpha_1=\alpha_2$	est $=\hat{\alpha}_1-\hat{\alpha}_2$, se$_0=\sqrt{\hat{\alpha}(\frac{1}{t_1}+\frac{1}{t_2})}$, $\hat{\alpha}=\frac{x_1+x_2}{t_1+t_2}$; HT: $z=\frac{\text{est}}{\text{se}_0}$
rate ratio, estimate and CI	$\hat{\phi}=\frac{\hat{\alpha}_1}{\hat{\alpha}_2}$; se$(\ln\hat{\phi})=\sqrt{\frac{1}{x_1}+\frac{1}{x_2}}$; 95% CI for $\ln\phi$: $\ln\hat{\phi} \pm 1.96\,\mathrm{se}(\ln\hat{\phi})$.
χ^2 ***goodness of fit test***	$u=\sum\frac{(o-e)^2}{e} \overset{\mathrm{d}}{\approx} \chi^2_{k-\ell}$ (provided $e{>}5$), where k= # classes, ℓ = # constraints
$r{\times}c$ contingency table	observed frequencies, $o=f_{ij}$
testing independence	expected frequencies $e=e_{ij}=\frac{f_{i\cdot}f_{\cdot j}}{n}$, where $f_{i\cdot}$ = row i sum, $f_{\cdot j}$ = col j sum
	$u=\sum\frac{(o-e)^2}{e} \overset{\mathrm{d}}{\approx} \chi^2_{(r-1)(c-1)}$ (provided $e{>}5$); for 2×2 table, $u \overset{\mathrm{d}}{\approx} \chi^2_1$.
2×2 contingency table	$\boxed{\begin{matrix} a & b \\ c & d \end{matrix}}$ $z=\frac{\text{est}}{\text{se}_0}=\frac{(ad-bc)\sqrt{a+b+c+d}}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$; $r=\frac{z}{\sqrt{n}}$, $u=z^2$.
odds ratio, estimate and CI	$\hat{\theta}=\frac{ad}{bc}$; se$(\ln\hat{\theta})=\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$; 95% CI for $\ln\theta$: $\ln\hat{\theta} \pm 1.96\,\mathrm{se}(\ln\hat{\theta})$.
Straight line regression	$Y_i \overset{\mathrm{d}}{=} \mathrm{N}(\alpha+\beta x_i,\sigma^2)$, $(i=1,2,\ldots,n)$.
least squares estimates	$\hat{\beta}=\frac{rs_y}{s_x} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$; $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$
estimate of σ^2	$s^2=\frac{1}{n-2}\Sigma(y-\hat{\alpha}-\hat{\beta}x_i)^2=\frac{n-1}{n-2}(1-r^2)s_y^2 = \frac{1}{n-2}\big(\Sigma(y-\bar{y})^2-\frac{(\Sigma(x-\bar{x})(y-\bar{y}))^2}{\Sigma(x-\bar{x})^2}\big)$
estimators	$\bar{y} \overset{\mathrm{d}}{=} \mathrm{N}(\alpha+\beta\bar{x},\frac{\sigma^2}{n})$, $\hat{\beta} \overset{\mathrm{d}}{=} \mathrm{N}(\beta,\frac{\sigma^2}{K})$, where $K=\Sigma(x-\bar{x})^2$; $\bar{y}$, $\hat{\beta}$ independent. $\hat{\mu}(x)=\bar{y}+(x-\bar{x})\hat{\beta} \overset{\mathrm{d}}{=} \mathrm{N}\big(\mu(x),c(x)\sigma^2\big)$, where $c(x)=\frac{1}{n}+\frac{(x-\bar{x})^2}{K}$
inference on β, $\hat{\mu}(x)$, $Y(x)$	$\hat{\beta} \overset{\mathrm{d}}{=} \mathrm{N}(\beta,\frac{\sigma^2}{K})$, $\hat{\mu}(x) \overset{\mathrm{d}}{=} \mathrm{N}\big(\mu(x),c(x)\sigma^2\big)$ $Y(x) \overset{\mathrm{d}}{=} \mathrm{N}\big(\mu(x),\sigma^2\big)$ $\frac{\hat{\beta}-\beta}{S/\sqrt{K}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-2}$; $\frac{\hat{\mu}(x)-\mu(x)}{S\sqrt{c(x)}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-2}$; $\frac{Y(x)-\hat{\mu}(x)}{S\sqrt{1+c(x)}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-2}$
CI for β, CI for $\mu(x)$, PI for $Y(x)$	$\hat{\beta} \pm c_{0.975}(\mathrm{t}_{n-2})\frac{s}{\sqrt{K}}$, $\hat{\mu}(x) \pm c_{0.975}(\mathrm{t}_{n-2})s\sqrt{c(x)}$, $\hat{\mu}(x) \pm c_{0.975}(\mathrm{t}_{n-2})s\sqrt{1+c(x)}$
Correlation	ρ $(-1 \leqslant \rho \leqslant 1)$ (population); r $(-1 \leqslant r \leqslant 1)$ (sample, estimate of ρ)
rank correlation, r'	correlation of ranks: $r'(x,y)=r(u,v)$ [critical values for r': Table 9]
distribution of r when $\rho=0$	$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-2}$ $[\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{\hat{\beta}}{S/\sqrt{K}}]$ $r=\frac{s_{xy}}{s_xs_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\ \Sigma(y-\bar{y})^2}}$
when $\rho \neq 0$	Statistic-Parameter diagram [Figure 10]

Probability Distributions		
1. Binomial distribution		$X \overset{\mathrm{d}}{=} \mathrm{Bi}(n,p)$ [n positive integer, $0 \leqslant p \leqslant 1$]
pmf, $p(x)$		$\binom{n}{x}p^xq^{n-x},\ x = 0, 1, 2, \dots, n;\ p+q = 1$ [Table 1]
physical interpretation		X = number of successes in n independent trials, each having probability p of success (Bernoulli trials)
$\mathrm{E}(X)$, $\mathrm{var}(X)$		$np,\ \ npq$
properties	(1)	If Z_i iidrvs $\overset{\mathrm{d}}{=} \mathrm{Bi}(1,p)$ then $X = Z_1 + Z_2 + \cdots + Z_n \overset{\mathrm{d}}{=} \mathrm{Bi}(n,p)$
	(2)	$X_1 \overset{\mathrm{d}}{=} \mathrm{Bi}(n_1,p), X_2 \overset{\mathrm{d}}{=} \mathrm{Bi}(n_2,p)$ indept $\Rightarrow X_1{+}X_2 \overset{\mathrm{d}}{=} \mathrm{Bi}(n_1{+}n_2,p)$
	(3)	If $n \to \infty, p \to 0$, so that $np \to \lambda$, then $\mathrm{Bi}(n,p) \to \mathrm{Pn}(\lambda)$
	(4)	If $n \to \infty$, then $\mathrm{Bi}(n,p) \sim \mathrm{N}(np, npq)$ [$np > 5, nq > 5$], in which case: if $X^* \overset{\mathrm{d}}{=} \mathrm{N}(np, npq)$, then $\Pr(X = k) \approx \Pr(k{-}0.5 < X^* < k{+}0.5)$ [CC]
2. Poisson distribution		$X \overset{\mathrm{d}}{=} \mathrm{Pn}(\lambda)$ [$\lambda > 0$]
pmf, $p(x)$		$\frac{e^{-\lambda}\lambda^x}{x!}$, $(x = 0, 1, 2, ...)$ [Table 3]
Poisson process		"events" occurring so that the probability that an "event" occurs in $(t, t+\delta t)$ is $\alpha\delta t + o(\delta t)$, where α = rate of the process
physical interpretation		X = number of "events" in unit time of a Poisson process with rate λ.
$\mathrm{E}(X)$, $\mathrm{var}(X)$		$\lambda,\ \ \lambda$
properties	(1)	$X_1 \overset{\mathrm{d}}{=} \mathrm{Pn}(\lambda_1), X_2 \overset{\mathrm{d}}{=} \mathrm{Pn}(\lambda_2)$ independent $\Rightarrow X_1 + X_2 \overset{\mathrm{d}}{=} \mathrm{Pn}(\lambda_1 + \lambda_2)$
	(2)	approximation to $\mathrm{Bi}(n,p)$ when n large, p small: $\lambda = np$.

	(3)	if $\lambda \to \infty$ then $\mathrm{Pn}(\lambda) \sim \mathrm{N}(\lambda, \lambda)$ [$\lambda > 10$], in which case:
		if $X^* \overset{\mathrm{d}}{=} \mathrm{N}(\lambda, \lambda)$, then $\Pr(X = k) \approx \Pr(k{-}0.5 < X^* < k{+}0.5)$ [CC]

3. Normal distribution		$X \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$ [$\sigma > 0$]
standard normal distribution		$\mathrm{N}(0,1)$
pdf, $\varphi(x)$; cdf, $\Phi(x)$		$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2};\quad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}dt$ [cdf: Table 5]
$\mathrm{E}(X)$, $\mathrm{var}(X)$, ν_3, ν_4		0, 1, 0, 3. [inverse cdf: Table 6]
general normal distribution, pdf, $f(x)$		$\dfrac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
physical interpretation		just about any variable obtained from a large number of components (by the central limit theorem)
$\mathrm{E}(X)$, $\mathrm{var}(X)$, ν_3, ν_4		$\mu,\ \sigma^2,\ 0,\ 3\sigma^4.$ (skewness = kurtosis = 0)
properties	(1)	if $X \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$ then $a + bX \overset{\mathrm{d}}{=} \mathrm{N}(a + b\mu, b^2\sigma^2)$
	(2)	$Z = \frac{X-\mu}{\sigma} \overset{\mathrm{d}}{=} \mathrm{N}(0,1) \Leftrightarrow X = \mu + \sigma Z \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2);\ \ c_q(X) = \mu + \sigma c_q(Z)$
	(3)	$X_1 \overset{\mathrm{d}}{=} \mathrm{N}(\mu_1, \sigma_1^2), X_2 \overset{\mathrm{d}}{=} \mathrm{N}(\mu_2, \sigma_2^2)$ indept $\Rightarrow X_1{+}X_2 \overset{\mathrm{d}}{=} \mathrm{N}(\mu_1{+}\mu_2, \sigma_1^2{+}\sigma_2^2)$
4. t distribution		$X \overset{\mathrm{d}}{=} \mathrm{t}_n$ [$n = 1, 2, 3, \dots$]
definition		if $Z \overset{\mathrm{d}}{=} \mathrm{N}(0,1), U \overset{\mathrm{d}}{=} \chi^2_n$ indept, then $X = \frac{Z}{\sqrt{U/n}} \overset{\mathrm{d}}{=} \mathrm{t}_n$
pdf, $f(x)$		$\frac{1}{\sqrt{n\pi}}\,\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\,\frac{1}{(1+\frac{x^2}{n})^{\frac{n+1}{2}}}\ \ (x > 0)$ [inverse cdf: Table 7]
$\mathrm{E}(X)$, $\mathrm{var}(X)$		$0,\ \frac{n}{n-2}$
comparison with standard normal		t_n has wider tails: var > 1; $\mathrm{t}_n \to \mathrm{N}(0,1)$ as $n{\to}\infty$: $(1{+}\frac{x^2}{n})^{-\frac{n+1}{2}} \to e^{-\frac{1}{2}x^2}$
5. χ^2 distribution		$X \overset{\mathrm{d}}{=} \chi^2_n$ [$n = 1, 2, 3, \dots$]
definition		if $Z_1, Z_2, \dots, Z_n$ iidrvs $\overset{\mathrm{d}}{=} \mathrm{N}(0,1)$ then $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \overset{\mathrm{d}}{=} \chi^2_n$
pdf, $f_X(x)$		$\frac{e^{-\frac{1}{2}x}x^{\frac{1}{2}n-1}}{2^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)}\ \ (x > 0)$ [inverse cdf: Table 8]
$\mathrm{E}(X)$, $\mathrm{var}(X)$		$n,\ 2n$
properties	(1)	$X_1 \overset{\mathrm{d}}{=} \chi^2_m, X_2 \overset{\mathrm{d}}{=} \chi^2_n$ indept $\Rightarrow X_1{+}X_2 \overset{\mathrm{d}}{=} \chi^2_{m+n}$
	(2)	sample on $\mathrm{N}(\mu, \sigma^2)$: $\frac{(n-1)S^2}{\sigma^2} \overset{\mathrm{d}}{=} \chi^2_{n-1} \Rightarrow \mathrm{E}(S^2) = \sigma^2, \mathrm{var}(S^2) = \frac{2\sigma^4}{n-1}$
	(3)	goodness of fit test: $\sum \frac{(o-e)^2}{e} \overset{\mathrm{d}}{=} \chi^2_{k-p-1}$