FURTHER PURE MATHEMATICS 4

(PROBABILITY AND STATISTICS)

~~Amy Khoo~~

Wan Lin

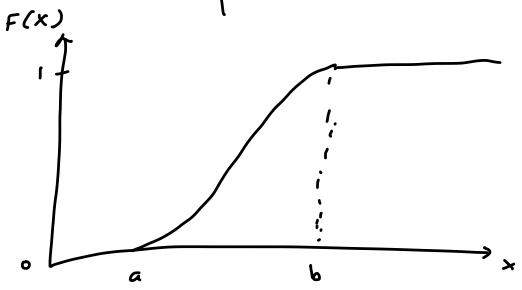
Chapter 1: Distributions

CDF

$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx.$

→ shortcut.

in the form: $F(x) = \begin{cases} 0 & x < a \\ g(x) & a \le x \le b \\ 1 & x > b \end{cases}$



For a pdf $f(x)$,

$$E(g(x)) = \int_{-\infty}^{\infty} f(x)\,g(x)\,dx$$

Transformation of RV

How do you find the pdf of $g(y)$ of the transformation $Y = g(X)$, w/ a given pdf $f(x)$?

eg $Y = X^2$
or $Y = 2X - 3$
or $Y = 5 - 2X$

methods:

- cdf technique
  - identify domain of X & Y
  - write $G(y) = P(Y \le y)$ (the CDF of y → the CDF of X)
  - $G'(y)$ → pdf of Y $(g(y))$.
- pdf technique (not needed).

e.g. 1 (Ex4Q2)

$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{10}(x^3 + x), & 0 \le x \le 2 \\ 1, & x > 2 \end{cases}$

Find the CDF of $A = X^3$.

Step 1: find CDF of $f(x)$.

$\therefore F(x) = \int_{-\infty}^{x} f(x)\,dx.$

Step 2: find $G(a)$, subst $A \to f(x)$

$G(a) = P(A < a)$
$= P(X^3 < a)$

Step 3: make $G(a)$ in the form $P(X < h(a))$, and solve accordingly using $F(x)$

$= P(X < \sqrt[3]{a})$
$G(a) = \frac{1}{10}\left((\sqrt[3]{a})^3 + \sqrt[3]{a}\right) = \frac{1}{10}(a + a^{\frac{1}{3}}).$

Step 4: create the CDF, taking the domain accordingly.

$\therefore G(a) = \begin{cases} 0, & a < 0 \\ \frac{1}{10}(a + \sqrt[3]{a}), & 0 \le a \le 8 \\ 1, & a > 8. \end{cases}$ ($0 \le x \le 2 \Rightarrow 0 \le x^3 \le 8$)

Step 5: you can find the PDF by differentiating accordingly.

$g(a) = \begin{cases} \frac{1}{10}(1 + \frac{1}{3}a^{-\frac{2}{3}}), & 0 \le a \le 8 \\ 0, & \text{otherwise} \end{cases}$

e.g.2 $f(x) = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & \text{otherwise.} \end{cases}$

Find pdf of Y, where $Y = 6X - 3$. → calc domain

$x = 0 \to y = -3$

$x = 2 \to y = 9.$

$F(x) \to \int_0^x \frac{1}{2}x \, dx$

$= \left[\frac{x^2}{4}\right]_0^x$

$= \frac{x^2}{4} - 0$

$= \frac{x^2}{4}$

$\therefore F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{4}, & 0 \le x \le 2 \\ 1, & x > 2. \end{cases}$

$G(y) = P(Y < y)$

$= P(6X - 3 < y)$

$= P\left(X < \frac{y+3}{6}\right)$

$= F\left(\frac{y+3}{6}\right)$

$\frac{\left(\frac{y+3}{6}\right)^2}{4} = \frac{(y+3)^2}{144}.$

$\therefore G(y) = \begin{cases} 0, & y < -3 \\ \frac{(y+3)^2}{144}, & -3 \le y \le 9 \\ 1, & y > 9 \end{cases}$

$\therefore g(y) = \begin{cases} \frac{(y+3)}{72}, & -3 \le y \le 9 \\ 0, & \text{otherwise.} \end{cases}$

e.g.3 $f(x) = \begin{cases} 3x^2, & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$

find pdf of Y, $Y = X^2$.

domain:

$x = 0 \to y = 0$

$x = 1 \to y = 1$

$F(x) = \int_0^x 3x^2 \, dx$

$= \left[x^3\right]_0^x$

$= x^3 - 0 = x^3.$

$\therefore F(x) = \begin{cases} 0, & x < 0 \\ x^3, & 0 < x < 1 \\ 1, & x > 1 \end{cases}$

$G(y) = P(Y < y)$

$= P(X^2 < y)$

$= P(X < \sqrt{y})$

$= F(\sqrt{y}) = y^{\frac{3}{2}}.$

$\therefore G(y) = \begin{cases} 0, & y < 0 \\ y^{\frac{3}{2}}, & 0 < y < 1 \\ 1, & y > 1. \end{cases}$

$\therefore g(y) = \begin{cases} \frac{3}{2}y^{\frac{1}{2}}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$

$0 < x < 1$: x^2 ↑ funct

e.g.4 $f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$

find pdf of Y. $Y = -\ln x$.

$x = 0 \to y = \infty$

$x = 1 \to y = 0$

$F(x) \to \int_0^x 1 \, dx$

$= [x]_0^x = x.$

$\therefore F(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < 1 \\ 1, & x \ge 1 \end{cases}$

$G(y) = P(Y < y)$

$= P(-\ln X < y)$

$= P(\ln X > -y)$

$= 1 - P(\ln X < -y)$

$= 1 - P(X < e^{-y}).$

$= 1 - F(e^{-y}).$

$= 1 - e^{-y}.$

$\therefore G(y) = \begin{cases} 0, & y \le 0 \\ 1 - e^{-y}, & y > 0 \end{cases}$

$\therefore g(y) = \begin{cases} e^{y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$

Chapter 2: Probability Generating Function

Fundamental definitions.

1) "generating funct"
→ a funct that stores data of a prob dist^n in a compact form.

2) "prob gen func (PGF)"
→ used only w/ disc prob dist^{n}s.
→ to obtain one, we introduce a "dummy variable" t, and the PGF will be a func of t.
→ formal def^n:
$G_X(t) = \sum_{\text{all } x} P(X=r)\,t^r$
$G_X(t) = E(t^X)$

Key results from PGF.

① $G_X(1) = 1$.
proof. $G_X(1) = \sum P(X=r)(1)^r$
$= \sum P(X=r)$
$= 1$.

② Power of t is the value of X.
ie $t^k \Leftrightarrow X=k$.

③ If $G_X(t)$ is expanded as a polynomial, the coeff of t^k gives $P(X=k)$.

④ Given the table

x	x_1	x_2	...	x_n
$P(X=x)$	p_1	p_2	...	p_n

it follows that
$G_X(t) = \sum_{k=1}^{n} p_k t^{x_k} = p_1 t^{x_1} + p_2 t^{x_2} \dots + p_n t^{x_n}$.

PGFs of Special Prob Dist^{n}s.

① Binomial, $X \sim B(n,p)$ $(n = 1, 2 \dots n)$

$G_X(t) = p_0 + p_1 t + p_2 t^2 + p_3 t^3 \dots + p_n t^n$
$= {}^nC_0(1-p)^n + {}^nC_1(p)(1-p)^{n-1} t + {}^nC_2(p)^2(1-p)^{n-2}t^2 \dots + {}^nC_n(p)^n t^n$

$\boxed{G_X(t) = (pt + (1-p))^n}$.

② Poisson, $X \sim Po(\lambda)$ $(n = 0, 1, 2 \dots \infty)$

$G_X(t) = p_0 + p_1 t + p_2 t^2 \dots + p_n t^n \dots$
$= \frac{e^{-\lambda}}{0!} + \frac{\lambda e^{-\lambda}}{1!} t + \frac{\lambda^2 e^{-\lambda}}{2!} t^2 \dots + \frac{\lambda^n e^{-\lambda}}{n!} t^n \dots$
$= e^{-\lambda}\left(1 + \frac{\lambda t}{1!} + \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} \dots \frac{\lambda^n t^n}{n!} \dots\right)$ → Taylor exp of $e^{\lambda t}$
$= e^{-\lambda}(e^{\lambda t})$

$\boxed{G_X(t) = e^{\lambda(t-1)}}$

③ Geometric, $X \sim Geo(p)$ $(n = 1, 2 \dots \infty)$

$G_X(t) = p_1 t^1 + p_2 t^2 + p_3 t^3 \dots + p_n t^n \dots$
$= pt + p(1-p)t^2 + p(1-p)^2 t^3 \dots + p(1-p)^{n-1} t^n \dots$
$= pt\,(1 + (1-p)t + (1-p)^2 t^2 \dots)$ $(a=1,\ r=(1-p)t)$
$= pt\left(\frac{a}{1-r}\right)$

$\boxed{G_X(t) = \frac{pt}{1-(1-p)t}}$

THE CONVOLUTION THEOREM

$G_{(X+Y)}(t) = G_X(t)\,G_Y(t)$,
if X & Y are indep.
⇒ the pgf of the Σ of indep rv is = to the product of their pgfs.
Extending this to three or more rv:
⇒ $G_{\Sigma X_k}(t) = \prod G_{X_k}(t)$

USING A PGF TO FIND μ & σ^2

- $E(X) = \sum x_k P(X=x_k)$
$= x_0 p_0 + x_1 p_1 \dots + x_2 p_2$.
- $E(X^2) = \sum x_k^2 P(X=x_k)$
- $Var(X) = E(X^2) - [E(X)]^2$.

we know $G_X(t) = \sum t^x P(X=x)$

$\therefore G'_X(t) = \sum x t^{x-1} P(X=x)$
$= \sum t^{x-1} \times P(X=x)$

let $t=1 \rightarrow G'_X(1) = \sum x P(X=x)$
$= E(X)$.

$\Rightarrow \boxed{G'_X(1) = E(X)}$.

$\therefore G''_X(t) = \sum x(x-1) t^{x-2} P(X=x)$
$= \sum (x^2 - x) P(X=x) t^{x-2}$
$= \sum x^2 P(X=x) t^{x-2} - \sum x P(X=x) t^{x-2}$

$\therefore G''_X(1) = \sum x^2 P(X=x) - \sum x P(X=x)$
$= E(X^2) - E(X)$.

$\therefore \boxed{E(X^2) = G''_X(1) + G'_X(1)}$.

$\therefore Var(X) = E(X^2) - [E(X)]^2$

or $\boxed{Var(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2}$.

* formulae in MF19.

LINEAR TRANSFORMATION OF VARIABLES *not stated in syllabus

Given a drv X, & $Y = aX + b$, we can find $G_Y(t)$.

$\Rightarrow G_Y(t) = E(t^Y)$
$= E(t^{aX+b})$
$= t^b E(t^{aX})$
$= t^b E((t^a)^X)$
$= t^b G_X(t^a)$

result: $\boxed{G_Y(t) = t^b G_X(t^a)}$. $(Y = aX + b)$

Chapter 3: Sampling, Confidence Interval and Hypothesis Testing

INTRODUCTION

- A "population" includes all the elements from a set of data.
- A "sample" consists of ≥ 1 observations from the population.
- "Sampling" involves extrapolating data from a sample of the population to the whole population.
- A "parameter" is a measurable characteristic of a population.
- A "statistic" is a measurable characteristic of a sample.

SAMPLING DIST^N OF SAMPLE MEAN

We use "$\bar{X}$" to denote the mean of a sample of a population.

$\Rightarrow$ this is a drv, with its' possible values being the mean of the different possible samples of the population.

e.g. for a population x_1, x_2, x_3, x_4, and sample size 2,

$\bar{X} \in \left\{ \frac{x_1+x_2}{2}, \frac{x_1+x_3}{2}, \frac{x_1+x_4}{2}, \ldots, \frac{x_3+x_4}{2} \right\}$.

① Mean of Sample Mean

The mean of the sampling dist^n of the mean will be equivalent to the mean of the population.

ie. $E(\bar{X}) = E(X) = \mu$.

② Variance of Sample Mean

The variance of the sampling dist^n of the mean is equivalent to the variance of the population divided by the sample size.

ie $Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\sigma^2}{n}$.

Central limit theorem (CLT)

The central limit theorem states that, given $n \geq 30$, any sampling dist^n tends towards $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

- Sampling Dist^{n}s of some special prob dist^{n}s

① Normal, $X \sim N(\mu, \sigma^2)$

$\Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

*don't forget the $(\frac{}{n})$!!

② Poisson, $X \sim Po(\lambda)$

$\Rightarrow \bar{X} \approx N(\lambda, \frac{\lambda}{n})$

③ Geometric, $X \sim Geo(p)$

$\Rightarrow \bar{X} \approx N(\frac{1}{p}, \frac{q^2}{pn})$

④ Binomial, $X \sim B(N, p)$

$\Rightarrow \bar{X} \approx N(Np, \frac{Np(1-p)}{n})$

☆ these are approximations. as we are estimating drv to a crv.

ESTIMATION

"Estimation" refers to the process in statistics where we can deduce characteristics about a population from measuring data from a sample.

⇒ this is done via the use of "sample statistics".

ESTIMATION METHODS

① Point estimate
- ⇒ a single value of a statistic.
- ⇒ eg $\bar{x}$ (sample mean) $\approx \mu$ (pop mean)

② Interval estimate
- ⇒ defined by 2 numbers.
- ⇒ eg $a < \mu < b$.

POINT ESTIMATION

Conceptualisation:
Let our sample contains n elements, $X_1, X_2 \dots X_n$ where $X_i \sim N(\mu, \sigma^2)$ for $1 \le i \le n$.

⇒ denote $\bar{X}$ as the sample mean, and

⇒ denote $\hat{\sigma}^2$ as the sample variance.

Keep in mind $\bar{X}$ and $\hat{\sigma}^2$ are also random variables, but are influenced by the values of X_i.

⇒ Hence, by def^n.

$$\bar{X} = \frac{\sum X_i}{n} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \left[\frac{\sum X_i}{n}\right]^2.$$

The unbiased estimate of the population mean

The value of $E(\bar{X})$, or $\hat{\mu}$, give us our estimate for the population mean.

⇒ from a sample we are making inferences about a population.

$$\Rightarrow E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}E(\sum X_i) = \frac{1}{n}E(X_1 + X_2 \dots + X_n) = \frac{1}{n} n E(X) = E(X) = \mu.$$

Conclusion := $\hat{\mu} = E(X) = \mu$

⇒ mean of sample mean is equal to mean of population.

$X \xrightarrow{\text{traditional}} E(X)$

$\bar{X} \longrightarrow E(\bar{X})$ alternative route (sampling)

The unbiased estimate of the population variance

Similarly, $E(\hat{\sigma}^2)$ will give us an estimate for the population variance.

$$\Rightarrow E(\hat{\sigma}^2) = E\left(\frac{\sum X_i^2}{n} - \left[\frac{\sum X_i}{n}\right]^2\right)$$
$$= \frac{1}{n}E(\sum X_i^2) - E(\bar{X}^2)$$
$$= \frac{1}{n}E(\sum X_i^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \frac{1}{n}E(X_1^2 + X_2^2 \dots X_n^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \frac{1}{n}E(nX^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= E(X^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \frac{1}{n}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)$$
$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2. \quad (\neq \sigma^2)$$

recall: $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - [E(\bar{X})]^2$
$\Rightarrow E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$.

recall: $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$
$\Rightarrow E(X^2) = \mathrm{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2$.

Conclusion := $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$.

★ the sample variance is not equal to the population variance!

⇒ How do we fix this?

Notice $\frac{n}{n-1}E(\hat{\sigma}^2) = \sigma^2$

$$E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \sigma^2.$$

⇒ hence, the unbiased estimate of the population variance is actually $\frac{n}{n-1}\hat{\sigma}^2$, which is also denoted as s^2.

ie $s^2 = \frac{n}{n-1}\hat{\sigma}^2$.

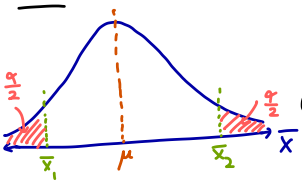
INTERVAL ESTIMATION (FOR μ ONLY)

Interval estimates are preferred over point estimates as it also provides a "confidence level" for the estimate.

⇒ these are called "confidence intervals".

eg 90% confidence interval ⇔ pop mean falls in interval.

Relationship bw pt est, confidence interval & z



From the diagram, it is apparent that:

① $P(\bar{x}_1 < \bar{X} < \bar{x}_2) = 1-\alpha$.

$\Rightarrow P(\bar{X} > \bar{x}_2) = P(\bar{X} < \bar{x}_1) = \frac{\alpha}{2}$.

② $\bar{x}_2 = -\bar{x}_1$.

Hence:

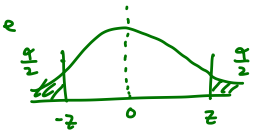
$$P\left(z > \frac{\bar{x}_2 - \bar{x}}{\hat{\sigma}}\right) = P\left(z < \frac{\bar{x}_1 - \bar{x}}{\hat{\sigma}}\right) = \frac{\alpha}{2}.$$

$\hat{\sigma} = \frac{\sigma}{\sqrt{n}}$

$$\Rightarrow P\left(z > \frac{\bar{x}_2 - \bar{x}}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right) = P\left(z < \frac{\bar{x}_1 - \bar{x}}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right) = \frac{\alpha}{2}.$$

Let $\frac{\bar{x}_2 - \bar{x}}{\left(\frac{\sigma}{\sqrt{n}}\right)} = z \Rightarrow \frac{\bar{x}_1 - \bar{x}}{\left(\frac{\sigma}{\sqrt{n}}\right)} = -z$.

ie



$\Rightarrow \bar{x}_2 = \bar{x} + z\frac{\sigma}{\sqrt{n}}$ $\quad \bar{x}_1 = \bar{x} - z\frac{\sigma}{\sqrt{n}}$.

Conclusion: The confidence interval is

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

(where $z = \phi^{-1}\left(\frac{\alpha+100}{200}\right)$ eg: $\alpha = 90\ (\%)$. $\therefore z = \phi^{-1}(0.95)$)

alternative ways to represent:

1) $\left(\mu - z\frac{\sigma}{\sqrt{n}},\ \mu + z\frac{\sigma}{\sqrt{n}}\right)$

2) $\mu \pm z\frac{\sigma}{\sqrt{n}}$

⇒ or: the true value of the population mean, μ, lies in this confidence interval.

☆ remember:

μ = pop mean

$\bar{x}$ = sample mean

$\hat{\mu}$ = mean of the sample mean dist^n

Cases

① μ & σ known

For $n \leq 30$, we make 2 assumptions:

- dist^n is normally distributed
- pop. variance is known

For $n \geq 30$, we apply CLT.

The confidence interval is just the direct application of the formula:

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

② μ is known, σ is unknown

We must approximate σ with s if we are not given σ directly, given that $n \leq 30$.

⇒ however, this introduces more uncertainty!

⇒ to compensate for this, instead of using a z-distribution, we use a t-dist^n instead for our calculations.

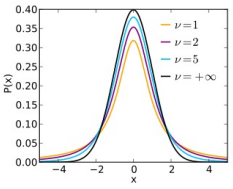
The confidence interval is given by

$$\bar{x} - t_{n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1}\frac{s}{\sqrt{n}}$$

t_{n-1} = critical value of t dist^n w/ $\nu = n-1$ & area of $\frac{\alpha}{2}$ in each tail.

t-DISTNS VS z-DISTNS

We use t-dist^{n}s to estimate a confidence interval when $n \leq 30$, σ is unknown, & $X \sim N(\mu, \sigma^2)$.



Characteristics

① Bell shaped and symmetrical but with fatter tails than the normal.

② Degree of freedom

⇒ denoted by ν

⇒ look at table in MF19

⇒ when $\nu \to \infty$, t dist^n → normal.

① If t dist^n is used:

$$z = \phi^{-1}_{\nu=?}\left(\frac{\alpha+100}{2}\right).$$

② If two / paired sample is used:

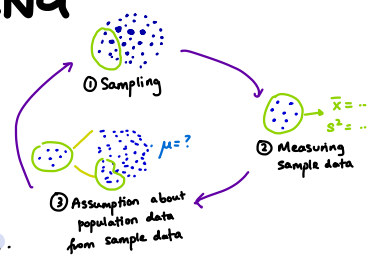
"$\frac{\sigma}{\sqrt{n}}$" = $s = \sqrt{s_c^2}$ or $\sqrt{s_p^2}$

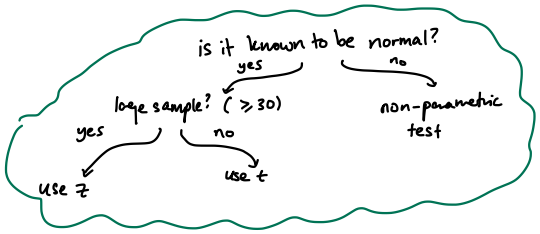
$$= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad = \sqrt{\frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{n_1+n_2-2}}$$

HYPOTHESIS TESTING

By examining characteristics (mean / variance) of a sample, we can make inferences about the population the sample was derived from.

↳ hypothesis testing can help us evaluate whether there is sufficient evidence to justify a given assumption about a population.





ONE SAMPLE z & t-TEST

An one sample test is used to determine whether a population parameter is significantly different than some hypothesised value.

z	t
↓	↓
$n \geq 30$	$n \leq 30$
OR	σ^2 unknown
$n \leq 30$	
$X \sim N$	
σ^2 known	

Steps

① State variable, check assumptions

② State the null (H_0) & alternative (H_1) hypothesis

H_1: $\mu > \mu_0$ "one-tailed", $\mu < \mu_0$ "one-tailed", $\mu \neq \mu_0$ "two-tailed"

*μ must be used, NOT $\bar{x}$.

Why? → assumption about population mean.

③ State sampling distⁿ according to H_0.

④ State level of test.

⑤ Find the critical value of the test, via the decision rule.

⑥ Perform the necessary calculations.

- z_t & z_c comparison
- $\bar{x}_t$ & $\bar{x}_c$ comparison
- p_{val} & α comparison.

⇒ write down the rejection criteria.

⑦ Form a conclusion.

(if H_0 rejected)

"We reject H_0 at (value of α) level of significance. There is sufficient evidence to conclude that (statement of the alternative hypothesis).

(and "H_0 accepted" statement follows same structure).

Worked example (z)

Example 6 Boys of a certain age are normally distributed and known to have a mean weight of $\mu = 85$ pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence, $n = 25$ boys (of the same age) are weighed and found to have a mean weight of 80.94 pounds. It is known that the population standard deviation σ is 11.6 pounds. Test at 5% significant level, what should be concluded concerning the complaint?

Step 1 (State variable of population)

Let X ↦ weight of boys/ pounds.

$\therefore X \sim N(\mu, 11.6^2)$

Step 2 (State hypothesises)

H_0 : $\mu = 85$

H_1 : $\mu < 85$.

Step 3 (State test & variable of sample, assuming H_0 is true)

→ z-test

$\therefore \bar{X} \sim N\left(85, \frac{11.6^2}{25}\right)$

Step 4 (State significance level)

$\alpha = 0.05$.

Step 5 & 6 (comparison & conditions for rejection) (choose 1 method)

1) Compare z_t to z_c (critical value)

$z_c = \Phi^{-1}(1 - 0.05)$

$= \Phi^{-1}(0.95)$

$= -1.645$.

★ We reject H_0 IF $z_t \leq -1.645$.

$z_t = \frac{\bar{x} - \hat{\mu}}{\hat{\sigma}}$ ($\bar{x}$ = test statistic of mean)

$= \frac{80.94 - 85}{\left(\frac{11.6}{\sqrt{25}}\right)}$ ($\hat{\mu}$ = expected mean)

$= -1.75$ (< -1.645) ⇒ can reject H_0!

Step 7 (State Conclusion)

"Hence, H_0 is rejected at 5% significance level. It subsequently follows that there is sufficient evidence to conclude the complaint that the boys living in the municipal children's home are underfed is valid."

Worked example (T)

Example 7 Monarch butterflies are bred at a butterfly farm. Monarch butterflies should grow to a have a mean wingspan of 9.4 cm. The breeders are concerned that their butterflies are growing as well as they could be, so a sample of six Monarch butterflies is taken and their wingspans (in cm), These are recorded as: 8.8, 9.6, 9.2, 9.1, 9.9, 8.7. Assuming that the wingspans are normally distributed and using a 10% significance level, investigate whether the wingspans of the butterflies are less than 9.4 cm.

Step 1 (variable)

Let X ↦ wingspan of butterflies / cm

⇒ $X \sim N(\mu, \sigma^2)$

Step 2 (hypothesis)

$H_0 \rightarrow \mu = 9.4$

$H_1 \rightarrow \mu < 9.4$

Step 3 (given H_0 true, find $\bar{X}$)

→ t-Test ∵ σ^2 unknown

$\bar{x} = \frac{\sum x}{n} = \frac{55.3}{6} = 9.21666$

$\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{510.75}{6} - 9.2166^2$

$= 0.178055$.

$\therefore s^2 = \frac{n}{n-1}\hat{\sigma}^2$

$= \frac{6}{5}(0.178055)$

$= 0.2136667$.

Step 4 (state α)

→ $\alpha = 0.1$

Step 5 & 6 (comparison)

$P(T \leq t_c) = 0.1$, $\nu = 5$

⇒ $P(T \leq -t_c) = 0.9$

$\therefore -t_c = 1.476$

$t_c = -1.476$. (crit value).

⇒ We reject H_0 if $t_t < -1.476$.

$t_t = \frac{\bar{x} - \hat{\mu}}{\left(\frac{s}{\sqrt{n}}\right)}$ ($\bar{x}$ = mean obtained from sample) ($\hat{\mu}$ = expected mean)

$= \frac{9.21666 - 9.4}{\sqrt{\frac{0.213667}{6}}}$

$= -0.9715$ (> -1.476).

Step 7 (conclusion)

From the results, we accept H_0 using a 10% significance level.

There is hence insufficient data to suggest the mean wingspan of the Monarch butterflies is less than 9.4cm.

TWO SAMPLE TEST

We perform a two-sample test on two random samples, each from 2 different populations.

⇒ For our syllabus, we test the difference between the 2 means.

Variance of the two samples

There are 2 ways in obtaining the variance of the two samples:

① Combined variance

case 1 variances are known

$$\Rightarrow \sigma_{comb} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

σ_1^2 = variance of pop. 1
σ_2^2 = variance of pop. 2
n_1 = sample size of 1
n_2 = sample size of 2.

case 2 variances are unknown

$$\Rightarrow S_{comb} \cong \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

★ If variances are the same, S is replaced by the S_p (pooled variance).

② Pooled Variance

We can calculate a pooled variance, otherwise known as the two-sample estimate of a common variance, when the means of the populations are different, but the variances can be assumed to be the same.

(MF19)

$$s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Alternate forms

$$\Rightarrow s^2 = \frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

$$\left(\frac{\Sigma(x-\bar{x})^2}{n} = \hat{\sigma}^2 = \frac{n}{n-1}s^2\right) \Rightarrow s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

DIFFERENCE BW TWO MEANS

We can use a two-sample test to test the significance of a difference between two population means.

Steps

① Define the variables.

$X_1 \sim N(\mu_1, \sigma^2)$ $X_2 \sim N(\mu_2, \sigma^2)$

② State the null and alternative hypothesises.

(one-tailed) $H_0: \mu_1 = \mu_2$ (two-tailed) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$ — case 1

or $\mu_1 < \mu_2$ — case 2

$H_1: \mu_1 \neq \mu_2$ — case 3

③ Compute the difference in the sample means.

$$d = \bar{x}_1 - \bar{x}_2 = \frac{\Sigma x_1}{n_1} - \frac{\Sigma x_2}{n_2}$$

④ Compute the combined variance.

i) If σ^2 is known:
⇒ compute directly.

$$\sigma_c^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \frac{(n_1+n_2)}{n_1 n_2}\sigma^2$$

*rmb variances of 1 & 2 are the same!

ii) If σ^2 is unknown:
⇒ compute pooled variance first,
⇒ then compute the combined variance. *estimate.

$$S_p^2 = \frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

$$S_c^2 = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} = \frac{(n_1+n_2)}{n_1 n_2}s_p^2$$

⑤ Using the given significance level, α, calculate critical values

i) If σ^2 is known, use the z-dist^n.

$z_c = \phi^{-1}(1-\alpha)$ (#1)
$= -\phi^{-1}(1-\alpha)$ (#2)
$= \pm\phi^{-1}(1-\frac{\alpha}{2})$ (#3)

ii) If σ^2 is unknown, use the t-dist^n, with $\nu = n_1 + n_2 - 2$.

$t_c = \phi^{-1}_{\nu=n_1+n_2-2}(1-\alpha)$ (#1)
$= -\phi^{-1}_{\nu=n_1+n_2-2}(1-\alpha)$ (#2)
$= \pm\phi^{-1}_{\nu=n_1+n_2-2}(1-\frac{\alpha}{2})$ (#3)

⑥ Using values from the samples, calculate the test values. (Assume H_0 is correct ⇒ $\mu_1 - \mu_2 = 0$)

i) If σ^2 is known

$$z_t = \frac{d-(0)}{\sigma_c}$$

$$z_t = \frac{d}{\sigma_c}$$

ii) If σ^2 is unknown

$$t_t = \frac{d-(0)}{s_c}$$

$$t_t = \frac{d}{s_c}$$

⑦ Check to see whether z_t / t_t lies within the rejection region determined by z_c / t_c.

You can then subsequently write your conclusion.

(lies in region) From the results, we can reject H_0 with a $(100\alpha)\%$ significance level.
It hence follows that there is sufficient evidence to suggest ..

(lies out of region) From the results, we can accept H_0 with a $(100\alpha)\%$ significance level.
It hence follows that there is insufficient evidence to suggest ..

Key Assumptions

- $X_1 \sim N(\,)$, $X_2 \sim N(\,)$
- values from the samples are independent from each other
- $\sigma_1^2 = \sigma_2^2$

PAIRED SAMPLE TEST

We can use a paired sample test when the two samples are correlated; for example, before & after, or when the samples occur in pairs.

★ z-test: $n \geq 30$ ★ t-test: $n < 30$

- objective: conclude the significance of the difference bw the two means.

Key assumptions

① For paired sample t-test, assume t-dist'n's.

② ★ Variance of the two samples are the same.

③ Treat the "difference" as a one-sample test.
(so like $\overline{X} \sim N(\hat{\mu}, \frac{\sigma^2}{n})$)

DIFFERENCE BW TWO MEANS IN PAIRED SAMPLE TEST.

① State variables, hypothesises & the significance level.
$H_0 : \mu_1 - \mu_2 = c$

② Tabulate the difference in the sample test means. $(\hat{\mu}_1 - \hat{\mu}_2)$.

③ Find t_c or z_c.

$t_c = \phi^{-1}_{\nu = n_1 + n_2 - 2}(\text{sig lvl})$ $\quad z_c = \phi^{-1}(\text{sig lvl})$

④ Find t_t or z_t.

$$t_t = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - c}{\left(\frac{s_d}{\sqrt{n}}\right)} \qquad z_t = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - c}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

⑤ Compare and make your conclusions.

* s_d = unbiased est. of variance of the diff of 2 samples.

- σ^2 same → $s_d = s_{pool}$
- σ^2 different → $s_d = s_{comb}$

Chapter 4: Chi-Square Tests

Chi-squared (χ^2) tests are primarily used for:

1. to test for the "goodness of fit" of an observed distⁿ to a theoretical one.
2. to test the null hypothesis that the variables are independent.

χ^2 STATISTIC

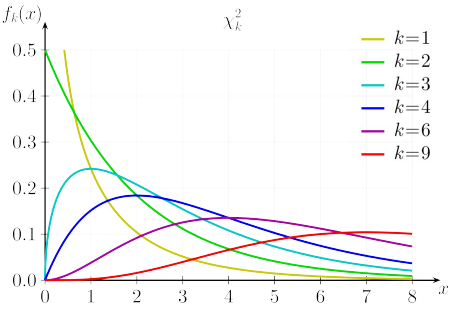
A χ^2 statistic is a measurement of how expectations compare to results.

* data used to calculate this MUST be:
  1) random;
  2) raw;
  3) mutually exclusive;
  4) drawn from independent variables; and
  5) drawn from a large enough sample.

CHI-SQUARE DISTⁿ

The χ^2-distⁿ is the distⁿ of the sum of squared std. normal deviates.

☆ std normal deviate = a random sample from the std. normal distⁿ.



- degrees of freedom, k, is equal to the number of std. normal deviates being summed.
  ⇒ this is also equal to the "level" of the variable.

χ^2 GOODNESS OF FIT TEST

We use the χ^2 goodness of fit test when:

→ sampling is purely random;
→ the variable is categorical; → can only take discrete values.
→ the expected value of the number of sample observations in each level of the variable is ≥ 5. → the different values the variable can take.

STEPS IN χ^2 GOF TEST.

1. State the hypothesises.
   H_0: data is consistent w/ a specified distⁿ.
   H_1: data is not consistent w/ a specifie distⁿ.
2. Formulate an analysis plan.
   The analysis plan describes how to use sample data to accept, or reject, H_0.
   a) State the significance level.
      → any value $\in (0,1)$
      → usually 0.01, 0.05, 0.10.
   b) Use the χ^2 GOF test.
      → does the observed sample frequencies differ significantly from the expected frequencies?
3. Analyse sample data.
   ⇒ expected freq of a value, $E_i = np_i$
   (n = total freq, p_i = prob of value)
   ⇒ χ^2 test statistic, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ (O_i = observed value, E_i = expected value.)
   ⇒ degrees of freedom, $\nu = k - 1 - \#$ of parameters est.
   (k = # of levels of categorical variable)
4. Calculate the critical value of χ^2.
   ☆ use the table in data booklet.
5. Make your conclusions.
   Compare χ^2_c and χ^2_t.
   If $\chi^2_t > \chi^2_c$ ⇒ reject H_0
   ↳ data not consistent w/ distⁿ.
   If $\chi^2_t < \chi^2_c$ ⇒ accept H_0.
   ↳ data consistent w/ distⁿ.

Worked example 1

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans but not All-Stars, and 10% are veteran All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

1. State hypothesises.
   H_0: consistent
   H_1: inconsistent.
2. Formulate analysis plan.
   $\alpha = 0.05$.
   χ^2 GOF Test.
3. Analyse sample data.

	O_i	$E_i = np_i$	$\frac{(O_i - E_i)^2}{E_i}$
rookies	50	100×0.3 = 30	13.33
veterans	45	100×0.6 = 60	3.75
all-stars	5	100×0.1 = 10	2.50
Σ	100	100	19.58

☆ if $E_i < 5$,
combine the row w/ the row below.
Repeat this until $E_i \geq 5$.

eg

E_i	E_i	E_i
0.12	0.46	5.06
0.34		
4.6	4.6	

$\therefore \chi^2_t = 19.58$.

4. Calculate critical value of χ^2.
   $\nu = 2$ $p = 0.95$
   (from the table) $\therefore \chi^2_c = 5.991$. ($< \chi^2_t$)
5. Conclusion.
   $\chi^2_t > \chi^2_c$
   ⇒ reject H_0
   ⇒ so proportion of cards proposed in the Q is wrong.

TEST FOR VARIABLE INDEPENDENCE USING A CONTIGENCY TABLE

In statistics, we can use a **contigency table** (also called a cross tabulation) as a way of displaying how **one categorical variable is "distributed" across another categorical variable**. (see right for example).

⇒ using χ^2 tests, we can determine whether any relationship is present between the variables (ie test for independence).

Example for further understanding

B \ A	a_1	a_2	a_3	total
b_1	100	200	100	400
b_2	350	400	250	1000
total	450	600	350	1400

The table shows how variable "A" is "distributed" across variable "B".

METHOD

Q: Given the following contigency table, determine whether a given person's income level affects their method of transport, using a 5% significance level.

		Method of Transport			
		Car	Public	Self	Total
Income Level	Small	58	21	36	115
	Average	199	49	64	312
	Large	205	32	29	266
	Total	462	102	129	693

① State hypothesises.

H_0 : income level & method of transport are independent

H_1 : " " " " not independent.

② For each entry in the contigency table, calculate the expected frequency.

Since we assume H_0 to be true, for a given entry "i" with row attr "A" & column attr "B", and for a d.r.v X, that takes a random entry in the table;

expected freq. $E_i = P(X=A) \times P(X=B) \times \Sigma f$

$= \frac{\text{row total}}{\text{grand total}} \times \frac{\text{column total}}{\text{grand total}} \times \text{grand total}$

$\Rightarrow E_i = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$.

E_i	Method of transport: Car	Public	Self
Small	76.666	16.926	21.407
Avg	208.000	45.922	58.077
Large	177.333	39.152	49.515

Income level

eg Small, car = $\frac{462 \times 115}{693} = 76.666..$

public, car = $\frac{102 \times 115}{693} = 16.926..$

self, car = $\frac{129 \times 115}{693} = 21.407..$

and so on.

③ For each entry "i", work out the squared std. normal deviate $\left(= \frac{(O_i - E_i)^2}{E_i}\right)$.

income level	method of transport	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
Small	car	58	76.666	5.5707
	public	21	16.926	0.9806
	self	36	21.407	9.9479
average	car	199	208.000	0.38942
	public	49	45.922	0.20630
	self	64	58.077	0.60406
large	car	205	177.333	4.3165
	public	32	39.152	1.3065
	self	29	49.515	8.4997

$\therefore \chi^2_t = 31.827.$

☆ same idea; if $E_i < 5$, combine the rows.

④ Analyse the data;
work out χ^2_c and χ^2_t, and their relationship. (< or >)

Degree of freedom = (# of rows − 1)(# of col. − 1)

(from ③) $\chi^2_t = 31.827.$

$\alpha = 0.95$
$\nu = (3-1)^2 = 4.$
$\therefore \chi^2_c = 15.51.$

$\Rightarrow \chi^2_t > \chi^2_c$.

⑤ Form appropriate conclusions.

⇒ Hence, we can reject H_0 with a 5% significance level.

⇒ Consequently, the method of transport and the income level are not independent.

Chapter 5: Non-Parametric Tests

Compared to parametric tests, which assume the data it is applied upon has a normal distⁿ, non-parametric tests do not assume anything about the data's underlying distⁿ.

→ usually, we use these tests when we know the population data does not have a normal distⁿ.

* Hence, a non-parametric test is any hypothesis test that do not rely on (any assumptions about) the shape of the distⁿ, or any population parameters (eg μ, σ^2 etc)

Why do we use non-parametric tests?

We generally use non-parametric tests when we know $X \sim N(\mu, \sigma^2)$;

however, there are other instances where we use them over their parametric counterparts:

1. The sample size is ↓↓↓;
2. The data contains irremovable outliers;
3. You want to test for the median rather than the mean;

Note: we generally do this for extremely skewed distns.

Why? ⇒ mean is affected by outliers.

TYPES OF NON-PARAMETRIC TESTS

Generally, non-parametric tests can be classified into several types:

1. Single Sample Test
   - one-sample sign test
   - one-sample Wilcoxon signed-rank test

   *replacement for 1-sample t test

   We use these to estimate the mean of a population, and compare it to a reference/target value.

   * main difference: Wilcoxon signed-rank test assumes a symmetric distⁿ.

2. Paired Sample Test
   - paired-sample sign test
   - paired-sample Wilcoxon signed-rank test

   *replacement for paired-sample t test

   We use these to test whether there exists a consistent difference between two interrelated populations.

   * again, main difference: Wilcoxon signed-rank test assumes a symmetric distⁿ.

3. Wilcoxon rank-sum test

   *replacement for 2-sample t-test

   We use the Wilcoxon rank-sum test to test whether there exists a consistent difference between two independent populations.

4. Chi-squared (χ^2) tests

   (recap) χ^2 tests are used to determine whether:

   (a) a given distⁿ fits a set of data; or

   (b) whether two groups of categorical data is independent (contigency table).

Pros and cons of non-parametric tests over parametric tests

- Advantages
  - ✓ fewer assumptions
  - ✓ caters for ↓↓ sample sizes
  - ✓ caters for more data types
- Disadvantages
  - ✗ less powerful than parametric tests (if assumptions are inviolated)
  - ✗ more tedious to calculate by hand

ONE SAMPLE TESTS

We use one sample non-parametric tests to compare the test median of a set of data against a hypothesised median value.

→ two types: { Sign test; and Wilcoxon signed-rank test.

1-SAMPLE SIGN TEST

Requirements:
1) each pair of data pts is ordered;
ie for $\{u_0, u_1, \dots u_k\}$. $u_i > u_{i-1} \quad \forall \ 1 \le i \le k$

Example: Let a sample obtained from a population be $\{1, 2, 2, 4, 5, 7, 10, 12, 15, 17\}$. Use a non-parametric test to test whether the median = 8, at a 5% significance level.

Method (one-tailed)

① State the hypothesises.

$H_0: m = m_0$

$H_1: m > m_0$ (case #1) or $m < m_0$ (case #2)

② (Case #1) Find S^+; (Case #2) Find S^-. } $= S$ (test statistic). * discard any elements = to the median. $(0_\downarrow \text{ or } 0_\uparrow = S)$.

③ If we consider H_0 to be true,

$P(x \le m_0) = P(x \ge m_0) = \frac{1}{2}$. (def^n of the median.)

Hence, if we let

$X \mapsto$ # of observations (case #1) $> m_0$, (case #2) $< m_0$, out of n,

we see $X \sim B(n, \frac{1}{2})$, as all observations are independent, and the chance any given observation will be larger than the median is equal to the chance the observation will be smaller than the median.

Given a significance level α, we say that we reject H_0 if

if $P(S^+ < s+) < \alpha$ (case #1)

or $P(S^- < s-) < \alpha$ (case #2).

Note #1 For 2-tailed tests, we check

$2P(X \ge S) \le \alpha$,

and if this is true we reject H_0.

#2 For large samples ($n \ge 30$), we can use a normal approximation;

$X \approx N(np, np(1-p))$.

Hence, $P(X \ge \text{or} \le S)$

$= P\left(Z \ge \text{or} \le \dfrac{S - \frac{n}{2}}{\sqrt{\frac{1}{4}n}}\right)$.

1-SAMPLE WILCOXON SIGNED-RANK TEST

Assumptions:
1) symmetric dist^n → mean = median.
2) continuous dist^n
3) data is independent.

Method

① State hypothesises.

H_0: pop. med = M

H_1: pop. med $\begin{cases} \ne M \\ > M \\ < M \end{cases}$.

② Calculate the test statistic:

a) Calculate the absolute differences bw each sample value & the hypothesised median.
ie $|x_1 - M|, |x_2 - M| \dots |x_n - M|$.

b) Rank each values from 1 to n, giving the lowest rank to the smallest absolute difference.

c) Calculate W^+ = the sum of ranks of the sample values $> M$;
& W^- = the sum of ranks of the sample values $< M$.

* ignore any values that are 0.

d) The test (signed rank) statistic, $x_t = \min(W^+, W^-)$.

③ Obtain the critical value, x_c, from MF19.
(this depends on: - the sample size, n (# of values that are not equal to m.) and - the level of significance, α.)

* value is different bw 1-tailed / 2-tailed - Refer to formula booklet.

④ Reject H_0 if $x_t \le x_c$.

Example

Wilcoxon ranked signed test
Example 4
The weights (in kg) of ten randomly selected Spanish mackerel are recorded:
1.6 1.1 2.1 2.4 2.2 2.9 2.6 2.3 2.7 1.9
Test, at the 5% significant level, whether the median weight is greater than 1.8 kg.

(1) H_0: median, $m = 1.8$
H_1: $m > 1.8$. } state hyp

(2)

W_i	$\lvert W_i - m \rvert$	sign	$W+$	$W-$
1.6	0.2	−		2
1.1	0.7	−		7
2.1	0.3	+	3	
2.4	0.6	+	6	
2.2	0.4	+	4	
2.9	1.1	+	10	
2.6	0.8	+	8	
2.3	0.5	+	5	
2.7	0.9	+	9	
1.9	0.1	+	1	
			$W^+ = 46$	$W^- = 9$

$\therefore x_t = \min(W^+, W^-)$
$= \min(46, 9)$
$= 9$.

(3) $n = 10$, $\alpha = 0.05$, 1-tailed
$\therefore x_c = 10$

(4) Since $x_t < x_c$, we reject H_0.
Hence, there is sufficient evidence to suggest the population median is greater than 1.8 kg.

Normal approximation

If we take many samples from the population, and record down the values of W^+, assuming H_0 is true,

⇒ W^+ can be modelled as a discrete random variable.

① Since the total of the ranks $= \frac{n(n+1)}{2}$, and (in the long run) it can be expected that $W^+ = W^- = \frac{1}{2}(\text{total})$,
hence $E(W^+) = \frac{1}{2}\left(\frac{n(n+1)}{2}\right) = \frac{n(n+1)}{4}$.

② $\text{Var}(W^+) = \frac{1}{24}n(n+1)(2n+1)$
(for now, take for granted. It is in MF19).

⇒ $W^+ \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$

⇒ Subsequently, $x_c = \phi^{-1}(\alpha)$ and reject H_0 if $x_t \le x_c$.
(see chapter 3).

W^+ or W^- (depending situ) $E(W)$ cc

$x_t = \dfrac{W - \mu_W \pm 0.5}{\sigma_W}$

$\sqrt{\text{Var}(W^+)}$

PAIRED SAMPLE TESTS

We use paired-sample tests to compare pairs of data from 2 different populations.

↳ specifically, we compare the differences between the data.

* alternative for paired-sample t test

PAIRED SAMPLE SIGN TEST

ASSUMPTIONS

The differences bw the pairs must:

1. be independent (from each other);
2. come from the same population dist^n (median is fixed).

* the populations themselves come from "pairs" of data (eg before & after); hence they are often not independent from each other.

METHOD

The methodology is similar to the 1-sample variant; however, the sign of the difference between the values is recorded instead.

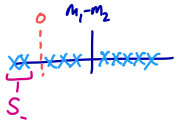
1. State hypothesises ($H_0: m_1 - m_2 = 0$; $H_1: m_1 - m_2 \begin{cases} > 0 \\ < 0 \\ \neq 0 \end{cases}$)
2. Calculate the sign of $(P_1 - P_2)$ for each pair. Record S_+ (# of "+" signs) and S_- (# of "−" signs).
   - ## once again, ignore any values = 0.
   - difference bw the two values in the pair.
3. S_+ or $S_- \sim B(n, \frac{1}{2})$.

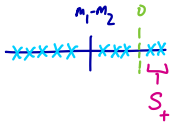
   Cases:

   a) If $H_1: m_1 - m_2 > 0$, ↳ $x_t = P(S_- \leq s_-)$;

   b) If $H_1: m_1 - m_2 < 0$, ↳ $x_t = P(S_+ \leq s_+)$;

   c) If $H_1: m_1 - m_2 \neq 0$, ↳ $x_t = 2P(S_+ \leq s_+)$.





4. Reject H_0 if $x_t \leq \alpha$. (α = sig level).

EXAMPLE

An investigation is carried out into the effectiveness of two types of post-operative pain relief drug: Drug1 and Drug2. Seven adults agree to take Drug1 on one day, and drug 2 on the second. The time, in hours, of pain relief is recorded.

	Drug1	Drug2
A	4.1	3.9
B	3.2	3.3
C	5.3	5.0
D	5.1	4.6
E	4.2	4.6
F	3.8	3.2
G	3.6	4.3

Ctest, at the 5% sig level, whether drug2 gives longer pain relief than drug1).

Let d = median of drug2 − median of drug1

(1) $H_0: d = 0$

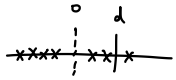
$H_1: d > 0$ $(m_2 > m_1)$

(2)

d2	d1	sign$(d_2 - d_1)$
3.9	4.1	−
3.3	3.2	+
5.0	5.3	−
4.6	5.1	−
4.6	4.2	+
3.2	3.8	−
4.3	3.6	+

$\Rightarrow S_+ = 3$, $S_- = 4$.

(3) $S_- \sim B(7, \frac{1}{2})$.



(a) $x_t = P(S_- \leq 4)$

$= (\frac{1}{2})^7 ({}^7C_4 + {}^7C_3 + {}^7C_2 + {}^7C_1 + {}^7C_0)$

$= 0.77343$.

(4) $0.77343 > 0.05$ $\therefore x_t > \alpha$.

Hence there is sufficient evidence to accept H_0.

Hence there is insufficient evidence to suggest that ... ...

WILCOXON MATCHED-PAIRS SIGNED-RANK TEST

ASSUMPTIONS

Again, this test has all the assumptions as the sign test, but with the additional criterion that

↳ the distribution of the differences must be symmetric.

METHOD

① State hypothesises. (see above)

② Calculate the sign of the differences $(P_1 - P_2)$, and the absolute value of the differences

③ Rank each difference in a similar manner to the tabulation outlined in the 1-sample test.

↳ Find W_+ and W_-.

④ Obtain $x_t = \min(W_+, W_-)$.

⑤ From MF19, get x_c. (using n - (# of values = 0), α and whether it is a 1 or 2-tailed test).

⑥ Reject H_0 if $x_t \leq x_c$.

NORMAL APPROXIMATION

Like the 1-sample test, we can use a normal approximation to estimate x_t and x_c for large values of n ($\frac{n(n+1)}{2} > 20$).

$\Rightarrow$ ① $\mu_W = \frac{n(n+1)}{4}$

② $\sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$.

$\Rightarrow$ $x_c = \phi^{-1}(q)$ and $x_t = \frac{W - \mu_W + 0.5}{\sigma_W}$,

(or $\phi^{-1}(1-\alpha)$)

and reject H_0 if $x_t < x_c$.

EXAMPLE (USING ABOVE Q)

(1) H_0 : $d = 0$ $(d = m_2 - m_1)$

H_1 : $d > 0$.

(2) (3)

d2	d1	$d_2 - d_1$	sign	W+	W-
3.9	4.1	0.2	−		2
3.3	3.2	0.1	+	1	
5.0	5.3	0.3	−		3
4.6	5.1	0.5	−		5
4.6	4.2	0.4	+	4	
3.2	3.8	0.6	−		6
4.3	3.6	0.7	+	7	
				$\therefore W_+ = 12$	$W_- = 16$.

(4) $x_t = \min(12, 16) = 12$.

(5) $n = 7$, $\alpha = 0.05$, 1-tailed $\therefore x_c = 3$.

(6) $x_t > x_c$

$\therefore$ accept H_0

$\therefore$ (conclusion).

WILCOXON RANK-SUM TEST

The Wilcoxon rank-sum test is the non-parametric alternative to the 2-sample t-test.

↳ unlike the paired sample tests, the two samples do not have to be the same size.

METHOD

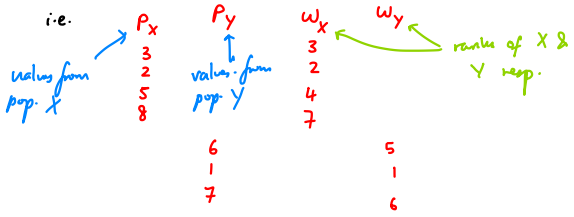
① State the hypothesises.

⇒ $H_0: m_1 - m_2 = 0$ (Let $d = m_1 - m_2$).

$H_1: m_1 - m_2 \begin{cases} > 0 \\ < 0 \\ \neq 0 \end{cases}$

② Rank each sample value, taking note of the population they are from.

i.e.



③ Calculate $W_X = \Sigma w_X$ & $W_Y = \Sigma w_Y$.

(In the above example, $W_X = 16$, $W_Y = 12$).

④ Calculate the sum of the ranks of the smallest sized sample when they are ranked the "other way around", $W_{X'}$. i.e. lowest rank = highest #.

Proof of formula. Since $W_X + W_{X'} = n_X(n_X + n_Y + 1)$;

(# of entries of n_X; sum of complementing ranks)

↳ it follows that

$$W_{X'} = m(n+m+1) - W_X.$$

⑤ Calculate the test statistic, x_t, where

$$x_t = \min(W_X, W_{X'})$$
$$(= \min(W_Y, W_{Y'}))$$
$$(= \min(W_X, n_X(n_X+n_Y+1) - W_X)).$$

⑥ Using:

- the sample sizes of the pop^{n}s n_X & n_Y, → *note: these are "m" and "n" in MF19, where $m \leq n$.
- the SL α, and
- whether the test is 1/2-tailed;

calculate x_t.

⑦ Reject H_0 if $x_t \leq x_c$.

EXAMPLE.

Researchers are investigating the effect of vitamin B12 on the size of the brain. A sample of males aged between 25 and 40 years is selected. Nine of them are known to have low B12 levels and seven are known to heave high levels. After a brain scan, the ratio of brain volume to skull capacity is recorded.

Low B12 levels	0.795	0.798	0.802	0.805	0.806	0.807	0.808	0.81	0.812
High B12 levels	0.786	0.789	0.792	0.796	0.799	0.8	0.803		

Carry out a Wilcoxon ran-sum test, at the 5% significant level, to see whether the level of vitamin B12 affects the size of the brain.

Let d = median of low B12 levels (X) − median of high B12 levels (Y).

(1) ⇒ $H_0: d = 0$

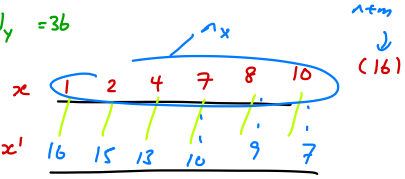
$H_1: d \neq 0$.

(2)

x	y	w_X	w_Y
~~0.795~~	~~0.786~~	4	1
~~0.798~~	~~0.789~~	6	2
~~0.802~~	~~0.792~~	9	3
0.805	~~0.796~~	11	5
0.806	~~0.799~~	12	7
0.807	~~0.800~~	13	8
0.808	~~0.803~~	14	10
0.810		15	
0.812		16	

* you can use a pencil to take note of the ranks crossed out.

(3) $n_X = 9$ $n_Y = 7$ $\therefore W_X = 100$ $W_Y = 36$



(4) $W_{Y'} = n_Y(n_X + n_Y + 1) - W_Y$ (in MF19.)

$= 7(9+7+1) - 36$

$= 83$.

$W_X + W_{X'} = (1+16) + (2+15) \ldots$

$= (n_X + n_Y + 1) n_X$

$W_{X'} = (n_X + n_Y + 1) n_X - W_X$

(5) $x_t = \min(W_Y, W_{Y'})$

$= \min(36, 83)$

$= 36$.

(6) $n_X = 9$ (= n), $n_Y = 7$ (= m), $\alpha = 0.05$, 2-tailed (for clarity)

$\therefore x_c = 40$.

(7) Since $x_t \leq x_c$, reject H_0.

(Make the relevant conclusions.)

NORMAL APPROXIMATION.

If n_X & n_Y (n, m) are large ($n_X, n_Y \geq 10$), we can approximate x_t with a normal distribution, with $E(x_t) = \frac{m(n+m+1)}{2}$

& $Var(x_t) = \frac{mn(m+n+1)}{12}$ (rmb n is the larger of n_X & n_Y.)

$$\Rightarrow x_t = \frac{x_t - \mu + 0.5}{\sigma} \quad \& \quad x_c = \phi^{-1}(\alpha).$$

and reject H_0 if $x_t \leq x_c$.