⑤

RANDOM FUNCTIONS ASSOCIATED W THE NORMAL DISTRIBUTION

$X_1, X_2 \dots X_n$ are independent random vbles with $\begin{cases} \mu_1, \mu_2 \dots \mu_n & \text{means} \\ \sigma_1^2, \sigma_2^2 \dots \sigma_n^2 & \text{variances} \end{cases}$ and

$Y = \sum_{i=1}^{n} c_i X_i \Rightarrow Y \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$

Ex: $X_1 \sim N(\mu=2, \sigma^2=3)$, $X_2 \sim N(\mu=1, \sigma^2=4)$ X_1, X_2 are indep.

Distribution of Y?

- $Y = 2X_1 + 3X_2 \Rightarrow Y \sim N(2\cdot 2 + 3\cdot 1, 2^2\cdot 3 + 3^2\cdot 4) = N \sim (7, 48)$
- $Y = X_1 - X_2 \Rightarrow Y \sim N(2-1, (1)^2(3)+(-1)^2(4)) = N \sim (1, 7)$

  $Y = X_1 - X_2 \Rightarrow Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

- $Y = X_1 + X_2 \Rightarrow Y \sim N(2+1, (1)^2\cdot(3) + (1)^2\cdot(4)) = N \sim (3, 7)$

  $Y = X_1 + X_2 \Rightarrow Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

σ^2 is the same!

SAMPLING DISTRIBUTION:

when $X_1, X_2 \dots X_n$ are observations from a random sample of n obs. from a normal dist. w mean μ and variance σ^2 $N(\mu, \sigma^2)$, then

- SAMPLE MEAN $\bar{X}$

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is normally distributed w $N\left(\mu, \frac{\sigma^2}{n}\right)$

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

the larger n, the smaller the variance $\frac{\sigma^2}{n}$

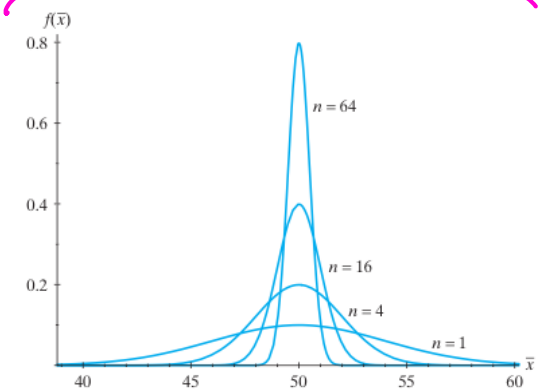


Figure 5.5-1 pdfs of means of samples from $N(50, 16)$

- SAMPLE VARIANCE S^2

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

- $\bar{X}$ and S^2 are independent

When sampling from a normal distribution:

Sum of squared differences of:

- population mean μ

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)} \qquad \frac{n(\bar{x}-\mu)}{\sigma^2} \sim \chi^2_{(1)}$$

- sample mean $\bar{X}$

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

we lose 1 df when we estimate the μ with $\bar{x}$

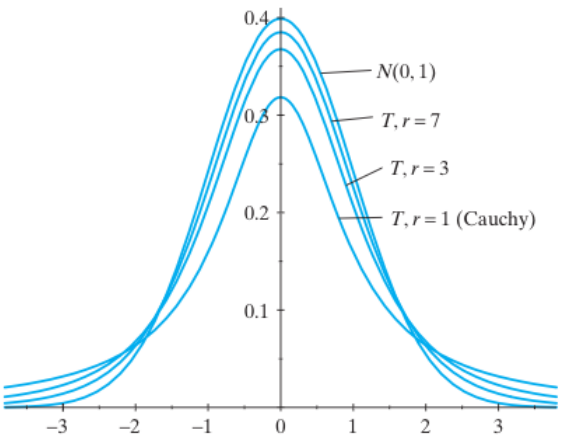
$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)} \Rightarrow \sum_{i=1}^{n}(x_i-\bar{x})^2 = s^2(n-1)$$

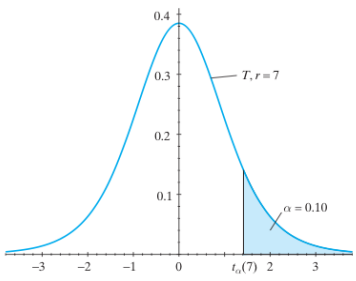
STUDENT'S t DISTRIBUTION:

$$T = \frac{Z}{\sqrt{U/r}},$$

where Z is a random variable that is $N(0,1)$, U is a random variable that is $\chi^2(r)$, and Z and U are independent. Then T has a t distribution with pdf

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty.$$



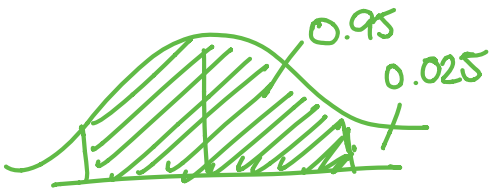


- support s is $-\infty < t < \infty$
- pdf { symmetric around 0
  bell-shaped
- density curve ~ standard normal curve but with "heavier" tails: it is more likely to get extreme values in the t dist.
- as the df r increase → t-dist approaches the N(0,1) dist

Ex.

$t_{(8)}$ $P(|T| < 2.306) = P(-2.306 < T < 2.306) =$

$P(T < 2.306) - P(T > 2.306) = 0.975 - 0.025 = 0.95$



-2.306 $\quad 2.306 = t_{0.025(8)}$

RANDOM VARIABLE T

Given a random sample $x_1, x_2 \dots x_n$ from a normal dist. $N(\mu, \sigma^2)$:

- $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- $U = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$
- Z and U are independent

$$T = \frac{Z}{\sqrt{\frac{U}{r}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{(n-1)}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{\sigma}{S} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

T has a Student's t dist with $(n-1)$ df.

OTHER DISTRIBUTIONS:

- GAMMA: $\gamma(\alpha, \theta)$

$Y \sim \gamma(\alpha = \frac{r}{2}, \theta = 2) \rightarrow Y \sim \chi^2_{(r)}$

If y follows a γ w $\alpha = \frac{r}{2}$ and $\theta = 2$ it follows a chi-square w r df.

- SUM OF INDEPENDENT χ^2 DISTRIBUTIONS $\sim \chi^2$ w sum of n df

$Y_i \sim \chi^2(r_i)$ w $i = 1, 2 \dots n \rightarrow V = \sum_{i=1}^{n} Y_i \sim \chi^2_{(\sum_{i=1}^{n} r_i)}$

- STANDARD NORMAL SQUARED $\sim \chi^2$ w 1 df

$Z \sim N(0,1) \rightarrow Z^2 \sim \chi^2_{(1)}$

- BETA DISTRIBUTION:

$X_1 \sim \gamma(\alpha, \theta, \beta) \quad X_2 \sim \gamma(\alpha, \theta, \beta)$

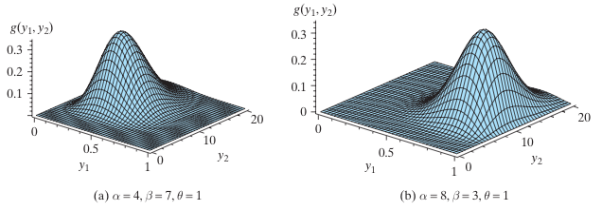
$X_1 = Y_1 Y_2 \qquad Y_1 = \frac{X_1}{X_1 + X_2}$

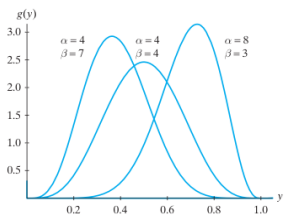
$X_2 = Y_2 - Y_1 Y_2 \qquad Y_2 = X_1 + X_2$

Marginal pdf of $Y_1 = f(y_1, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot y_1^{\alpha-1} \cdot (1-y_1)^{\beta-1}$

$Y_1 \sim \beta(\alpha, \beta)$

$0 < y_1 < 1$

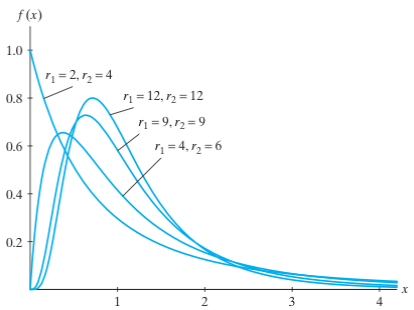




- F DISTRIBUTION: w r_1 and r_2 df.

$V_1 \sim \chi^2_{(r_1)} \qquad V_2 \sim \chi^2_{(r_2)}$

$W = \frac{\frac{V_1}{r_1}}{\frac{V_2}{r_2}} \qquad W \sim F_{(r_1, r_2)}$



$$\text{pdf } f(w) = \frac{(r_1/r_2)^{r_1/2}\,\Gamma[(r_1+r_2)/2]\,w^{r_1/2-1}}{\Gamma(r_1/2)\,\Gamma(r_2/2)\,[1+(r_1 w/r_2)]^{(r_1+r_2)/2}}$$

CENTRAL LIMIT THEOREM

The sampling distribution is approximately normally distributed regardless of the distribution of the underlying random sample when n is sufficiently large.

Random sample $x_1, x_2 \dots x_n$ from any distribution w/ μ and σ^2 w sufficiently large n:

- sample mean $\bar{X} \sim N\left(E(\bar{x}) = \mu_{\bar{x}} = \mu,\ Var(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\right)$

$$\bar{x} \xrightarrow[n\to\infty]{} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{} N(0,1)$$

APPROXIMATIONS FOR DISCRETE DISTRIBUTIONS

NORMAL APPROXIMATION TO BINOMIAL

$y \sim b(n,p)$ is the sum of n $x_i \sim$ Bernoilli (p) random variables. $x_1, x_2 \dots x_n$

$$y = \sum_{i=1}^{n} x_i$$

	mean	variance
$x_i \sim$ Bernoilli (p) with	$\mu = Ex = p$	$\sigma^2 = Var(x) = p(1-p) = pq$
$y \sim$ binomial (n,p) with	$\mu = np$	$\sigma^2 = np(1-p) = npq$

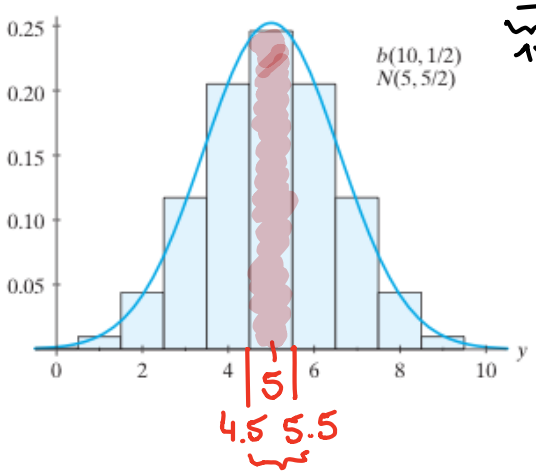
CLT: $y \sim N(np, npq)$ if n is sufficiently large $\begin{cases} np \geq 5 \\ nq \geq 5 \end{cases}$

$$w = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{x}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \quad n\to\infty$$

$$w = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sqrt{n}\sigma} = \frac{y-np}{\sqrt{n}\cdot\sqrt{pq}} = \frac{y-np}{\sqrt{npq}} \sim N(0,1) \quad n\to\infty$$

CONTINUITY CORRECTION

We approximate probabilities for discrete distributions with a continuous distribution. $P(Y=k)$ is represented by the area of the rectangle



→ height = $P(Y=k)$

base = length = 1 centered around k

$P(Y=k) = P(k-1/2 < Y < k+1/2)$

$P(Y=5) = P(4.5 < Y < 5.5)$

$= P\left(\frac{4.5-5}{\sqrt{2.5}} < Z < \frac{5.5-5}{\sqrt{2.5}}\right) = P(-0.32 < Z < 0.32)$

$= P(Z<0.32) - P(Z>0.32) = 0.625 - 0.3745 =$

$= 0.251$

SAMPLE PROPORTION : $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the proportion of the sample that meets condition of interest.

It can be estimated by:

$$Z = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{\frac{\sum_{i=1}^{n} X_i - np}{n}}{\sqrt{\frac{np(1-p)}{n^2}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{n\to\infty} N(0,1)$$

NORMAL APPROXIMATION TO POISSON

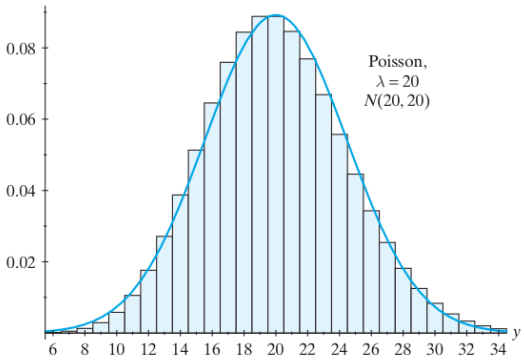
$X_1, X_2 \dots X_n$ are independent Poisson random vbles

$X_i \sim Pois(\lambda)$ with mean = 1

$Y = \sum_{i=1}^{n} X_i$ $Y \sim Pois(\lambda)$ with mean = λ and variance = λ

CLT: $Y \sim Poisson(\lambda)$

$W = \frac{Y-\lambda}{\sqrt{\lambda}} \xrightarrow{n\to\infty} N(0,1)$ when λ is sufficiently large


0.08
0.06
0.04
0.02
Poisson,
λ = 20
N(20, 20)
6 8 10 12 14 16 18 20 22 24 26 28 30 32 34
y

ESTIMATION

Random sample:

observed values of a specific random sample: $x_1, x_2 \dots x_n$ lowercase

Random variables arising from a random sample: $X_1, X_2 \dots X_n$ uppercase

- PARAMETER: θ
- PARAMETER SPACE: Ω range of possible values of parameter θ

Ex: proportion of students who smoke $\Omega = \{p : 0 \leq p \leq 1\}$

- STATISTIC: the function of $X_1, X_2 \dots X_n$ used to estimate parameter θ

Ex: the statistic of relative frequency $\frac{\text{n° of succes}}{n}$ estimates p.

- POINT ESTIMATOR: the statistic $u(X_1, X_2 \dots X_n)$ used to estimate θ

★ diff between point est and statistic

Ex: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a point estimator of population μ

$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ " "population proportion p

- POINT ESTIMATE: the function of $u(x_1, x_2 \dots x_n)$ computed from a set of data is an observed point estimate of θ

Ex: x_i = observed grade point avg. of $n = 88$ students

$\bar{x} = \frac{1}{88}\sum_{i=1}^{88} x_i = 3.12$ is a point estimate of the pop. μ.

- ESTIMAND: what is being estimated, often a parameter θ

Ex: μ.

MAXIMUM LIKELIHOOD ESTIMATION

what for?

Random sample $X_1, X_2 \dots X_n$ — ex $X_1, X_2 \dots X_n$

with a prob. distribution that depends on parameter θ → $X_i \sim N(\mu, \sigma^2)$

Goal: find a point estimator $u(X_1, X_2 \dots X_n)$ → $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ to estimate the μ

such that $u(x_1, x_2 \dots x_n)$ is a "good" point estimate

the observed values of a random sample

LIKELIHOOD FUNCTION: $L(\theta)$ (for 1 parameter)

Random sample: $X_1, X_2 \dots X_n$. X_i is independent bc it's random sample.

the pdf for each X_i is $f(x_i, \theta)$

Joint probability mass function of X_i independent vbles

$$L(\theta) = P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = f(x_1; \theta) \cdot f(x_2; \theta) \dots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Ex: Random sample $X_1, X_2 \dots X_n$. $X_i = 0$ → NO car; $X_i = 1$ → car

$X_i \sim$ Bernoulli(p) → X_i are independent Bernoulli random vbles

pmf of each X_i: $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$ for $x_i = 0$ or 1, $0 < p < 1$

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \dots p^{x_n}(1-p)^{1-x_n} =$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

Find the p that maximizes the likelihood of $L(p)$

↳ differentiate the $L(p)$ (to find where the function has a maximum → review calculus)

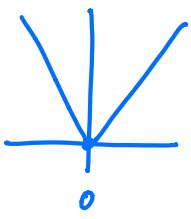
$\frac{d}{dp} L(p) = L'(p) = 0$

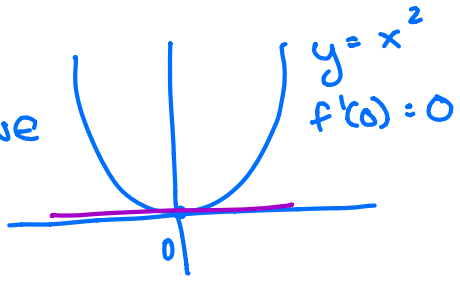
CALCULUS REVIEW :

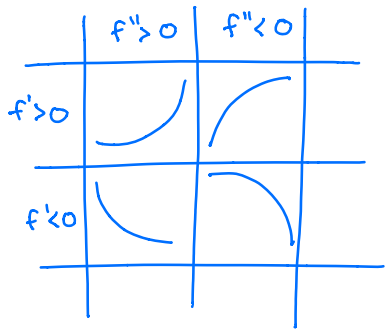
EXTREMUMS :

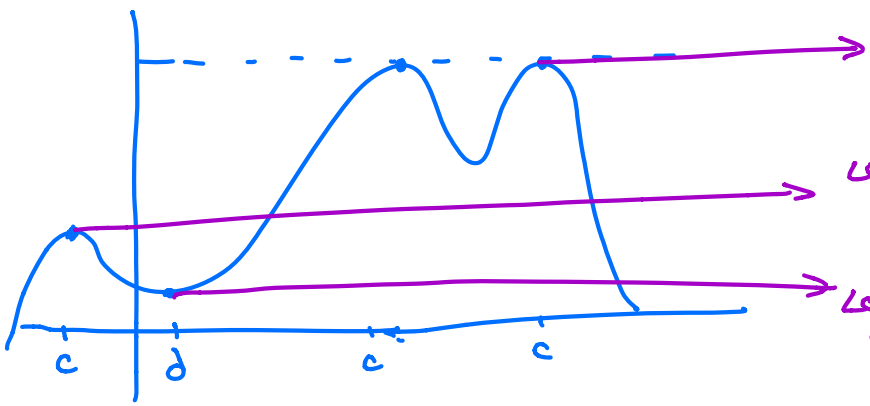
FERMAT'S THEOREM : suppose f is a fnction defined in the interval $(a,b) \ni c \Rightarrow c$ is in (a,b). if $f(c)$ is differentiable at c:

- $f(c) \neq 0 \Rightarrow c$ is not a extreme value of f.
- $f'(c) = 0 \Rightarrow c$ is a extreme value of f = local extremum.
  c is a critical point where the tangent is horizontal
- $f'(c) = DNE \Rightarrow c$ is a extreme value of f = local extremum.









GLOBAL MAXIMUM : for any x in domain of f
$f(c) \geq f(x)$

LOCAL MAXIMUM : ϵ = small number > 0
for any x in $(c-\epsilon, c+\epsilon)$ = "near c"
$f(c) \geq f(x)$

LOCAL MINIMUM : for any x in $(d-\epsilon, d+\epsilon)$ $f(d) \leq f(x)$

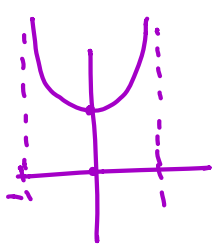
EXTREME VALUE THEOREM : if $f(x)$ is continuous in the closed interval $[a,b]$ then there are c and d in $[a,b]$ so that for all x in $[a,b]$
$f(x) \leq f(c) \Rightarrow f(x)$ attains a MAXIMUM VALUE
$f(d) \leq f(x) \Rightarrow f(x)$ attains a MINIMUM VALUE

HOW TO FIND MAX MIN VALUES:

1. Differentiate
2. List critical points
3. Check them
4. Check limiting behaviour

ex: $f(x) = \frac{1}{(x^2-1)^2}$ on $(-1,1)$

1. $f'(x) = -2(x^2-1)^{-3} \cdot 2x = \frac{-4x}{(x^2-1)^3} = 0$ DNE
2. $-4x = 0 \Rightarrow x = 0$
3. $f'(x)$: -1 (−) 0 (+) 1
4. $\lim_{x \to -1^+} f(x) = \infty$
   $\lim_{x \to 1^+} f(x) = \infty$



CALCULUS REVIEW CONT'

ex: $f(x) = x - |x^2 - 2x|$ in $[0, 3]$

$$f(x) = \begin{cases} x - (x^2 - 2x) & \text{if } x^2 - 2x \geq 0 \\ x + (x^2 - 2x) & \text{if } x^2 - 2x < 0 \end{cases}$$

① $f'(x)$

① $f'(x) = 1 - (2x - 2) = 3 - 2x$ if $x^2 - 2x > 0 \Rightarrow x(x-2) > 0$ → $x > 2$ or $x < 0$

② $f'(x) = 1 + (2x - 2) = -1 + 2x$ if $x^2 - 2x < 0 \Rightarrow x(x-2) < 0$ → $x < 2$ or $x > 0$

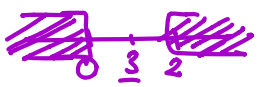
$x = 2$: $f'(2)^{①} = 3 - 4 = -1$, $f'(2)^{②} = 3$ ⇒ $\neq$

$x = 0$: $f'(0)^{①} = 3$, $f'(0)^{②} = -1$ ⇒ $\neq$

$f'(x) = DNE$ not differentiable

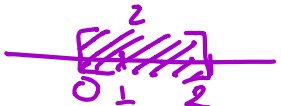
② $f'(x) = 0$

① $3 - 2x = 0 \Rightarrow x = \frac{-3}{-2} = \frac{3}{2}$



not in interval

② $-1 + 2x = 0 \Rightarrow x = \frac{1}{2}$



in interval

$\boxed{x = \frac{1}{2}}$

Do derivative without taking log:

$$L'(p) = \frac{d}{dp}\, p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i} = \text{product rule } \frac{d}{dx} f(x)\cdot g(x) = f'(x)\cdot g(x) + g'(x) f(x)$$

chain rule $= f'(g(x))\cdot g'(x)$

$f(g(x)) = (1-p)^{n-\sum x_i}$

$f'(g(x)) = (n-\sum x_i)(1-p)^{n-\sum x_i - 1}$

$g(x) = (1-p)$

$g'(x) = -1$

$$= \frac{d}{dp} p^{\sum x_i} \cdot (1-p)^{n-\sum x_i} + p^{\sum x_i} \cdot \boxed{\frac{d}{dp}(1-p)^{n-\sum x_i}}$$

$$= \sum_{i=1}^{n} x_i\, p^{\sum x_i - 1} \cdot (1-p)^{n-\sum x_i} + p^{\sum x_i} \cdot \underbrace{(n-\sum_{i=1}^{n} x_i)(1-p)^{n-\sum x_i - 1}}_{f'(g(x))} \cdot \underbrace{(-1)}_{g'(x)} =$$

$$= \sum_{i=1}^{n} x_i\, p^{\sum x_i - 1} \cdot (1-p)^{n-\sum x_i} - p^{\sum x_i} \cdot (n-\sum_{i=1}^{n} x_i)(1-p)^{n-\sum x_i - 1} = 0 \qquad 0<p<1$$

$$= \sum_{i=1}^{n} x_i \cdot p^{\sum x_i - 1} \cdot p^{-1} \cdot (1-p)^{n-\sum x_i} - p^{\sum x_i} \cdot (n-\sum x_i)\cdot (1-p)^{n-\sum x_i} \cdot (1-p)^{-1} =$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}\left[\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}\right] = 0 \quad \text{this} = 0 \text{ when}$$

$$\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0 \Rightarrow \frac{\sum x_i}{p} = \frac{n-\sum x_i}{1-p} \Rightarrow \sum x_i (1-p) = (n-\sum x_i)p$$

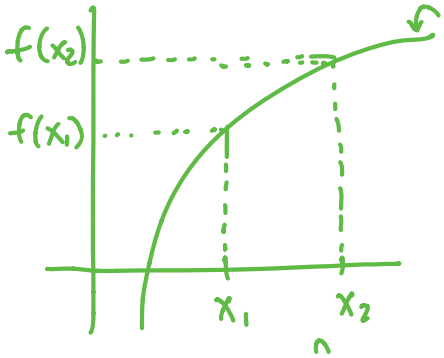
$$= \sum x_i - \cancel{\sum x_i p} = np - \cancel{\sum x_i p}$$

$$\sum_{i=1}^{n} x_i - np = 0 \Rightarrow p = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$L''(\bar{x}) < 0$ so $L(\bar{x})$ is a maximum

The statistic $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$ is the MAXIMUM LIKELIHOOD ESTIMATOR.

To make differentiation easier take the logaritum of the $f(x_i;p)$



↑ $y = \ln(x)$ the natural logaritum is an increasing fnction of x

$\Rightarrow$ if $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$

so the value of p that maximizes the logaritum of the likelihood function $\ln(L(p))$ also maximizes the likelihood fnction $L(p)$

$$L(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

CHAIN RULE

$$\ln(L(p)) = \sum_{i=1}^{n} x_i \cdot \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \boxed{\ln(1-p)} \qquad \frac{\partial}{\partial p} \ln(1-p) = \frac{1}{1-p} \cdot (-1)$$

$$\frac{\partial \ln(L(p))}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} = 0 \Rightarrow \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \cdot p(1-p) = 0$$

$$\sum x_i (1-p) - (n - \sum x_i) \cdot p = 0 \Rightarrow \sum x_i - \cancel{\sum x_i p} - np + \cancel{\sum x_i p} = 0$$

$$\sum x_i - np = 0 \Rightarrow \hat{p} = -\frac{\sum x_i}{-n} = \frac{\sum x_i}{n} \rightarrow \text{estimate (lowercase x)}$$

$$\hat{P} = \frac{\sum X_i}{n} \rightarrow \text{estimator (uppercase X)}$$

To verify we obtained a maximum $\rightarrow f''(\ln(p)) < 0$

Given a random sample $X_1, X_2 \dots X_n$ from a distribution that depends on $\theta_1, \theta_2 \dots \theta_m$ unknown parameters with $\begin{cases} pmf\ f(x_i; \theta_1, \theta_2 \dots \theta_m) \\ \text{restricted to parameter space} \curvearrowright \Omega \end{cases}$ then:

- LIKELIHOOD FUNCTION: $L(\theta_1, \theta_2 \dots \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2 \dots \theta_m)$
- MAXIMUM LIKELIHOOD ESTIMATOR: $\hat{\Theta}_i = u_i(X_1, X_2 \dots X_n)$ of θ_i for $i = 1, 2 \dots m$
  if $[u_1(x_1, x_2 \dots x_n), u_2(x_1, x_2 \dots x_n) \dots u_m(x_1, x_2 \dots x_n)]$ is the m-tuple that maximizes $L(\theta_1, \theta_2 \dots \theta_m)$ (the likelihood fnction)
- MAXIMUM LIKELIHOOD ESTIMATES: the corresponding observed values of the statistics $[u_1(x_1, x_2 \dots x_n), u_2(x_1, x_2 \dots x_n) \dots u_m(x_1, x_2 \dots x_n)]$ of θ_i for $i = 1, 2 \dots m$

For a given MLE / statistic $u(X_1, X_2 \dots X_n)$,

$u(X_1, X_2 \dots X_n)$ is an

if $E(u(X_1, X_2 \dots X_n)) = \theta \rightarrow$ UNBIASED ESTIMATOR of parameter θ

if $E(u(X_1, X_2 \dots X_n)) \neq \theta \rightarrow$ BIASED ESTIMATOR of parameter θ

Ex. $X_i \sim$ Bernoulli (p). Is the MLE of $p \rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a [biased / unbiased] estimator?

$E(\hat{p}) = p$

$E(X_i) = p$

$E(\hat{p}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n} \cdot np = p \rightarrow$ UNBIASED!

Ex: $X_i \sim N(\mu, \sigma^2)$ MLE $\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$, $\hat{\sigma}_2 = \frac{\sum (x_i - \bar{x})^2}{n}$ are they [biased / unbiased] estimators?

$E(X_i) = \mu$

$E(\hat{\mu}) = E\left(\frac{\sum x_i}{n}\right) = \frac{1}{n} E(\sum x_i) = \frac{1}{n} \sum E(x_i) = \frac{1}{n} \sum \mu = \frac{1}{n} n\mu = \mu$ UNBIASED!

$E(\hat{\sigma}^2) = ?$

$\hat{\sigma}_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) =$

$= \frac{1}{n}\sum x_i^2 - 2\bar{x} \cdot \underbrace{\frac{1}{n}\sum x_i}_{\bar{x}} + \frac{1}{n}\sum \bar{x}^2 = \frac{1}{n}\sum x_i^2 - 2\bar{x}^2 + \frac{1}{n} \cdot n\bar{x}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 \Rightarrow$

$E(\hat{\sigma}^2) = E\left[\left(\frac{1}{n}\sum x_i^2\right) - \bar{x}^2\right] = E\left(\frac{1}{n}\sum x_i^2\right) - E(\bar{x}^2) = \frac{1}{n}\sum E(x_i^2) - E(\bar{x}^2) =$

$Var(X) = \sigma^2 = E(X^2) - E(X)^2 = E(X^2) - \mu^2 \Rightarrow E(X^2) = \sigma^2 + \mu^2$

$Var(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{x}^2) - \mu^2$

$= \frac{1}{n} \sum (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{1}{n}(n\sigma^2 + n\mu^2) - \frac{\sigma^2}{n} - \mu^2 = \sigma^2 - \frac{\sigma^2}{n} + \mu^2 - \mu^2 =$

$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n} = \frac{\sigma^2(n-1)}{n} \neq \frac{\sum (x_i - \bar{x})^2}{n}$ BIASED ESTIMATOR!

Is S^2 an unbiased estimator ?

Recall: sample variance of $X_i \sim N(\mu, \sigma^2)$: $S^2 = \frac{1}{n-1} \cdot \hat{\sum}_{i=1} (X_i - \bar{X})^2 \Rightarrow \hat{\sum}_{i=1} (X_i - \bar{X}) \cdot S^2 (n-1)$

- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$
- $X \sim \chi^2_{(r)} \rightarrow E(X) = r \Rightarrow$ the df.

$$E(S^2) = E\left(\frac{\hat{\sum}(X_i - \bar{X})^2}{n-1}\right) = E\left(\frac{S^2(n-1)}{(n-1)} \cdot \frac{\sigma^2}{\sigma^2}\right) = E\left(\frac{S^2(n-1)}{\sigma^2} \cdot \frac{\sigma^2}{(n-1)}\right) =$$

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

S^2 is UNBIASED of σ^2 ESTIMATOR !!

$\frac{S^2(n-1)}{\sigma^2}$: $X_i \sim \chi^2_{(n-1)}$, $E(X_i) = (n-1)$

S^2 is ALWAYS an unbiased estimator of σ^2, but not of σ.

METHOD OF MOMENTS

for $K = 1, 2, \ldots$

K^{th}	THEORETICAL MOMENT	SAMPLE MOMENT
About the origin	$E(X^K)$	$M_K = \frac{1}{n} \hat{\sum}_{i=1} X_i^K$
About the mean	$E[(X-\mu)^K]$	$M_K = \frac{1}{n} \hat{\sum}_{i=1} (X_i - \bar{X})^K$

- ABOUT THE ORIGIN:

① Equate theoretical moments with sample moments

- 1^{st}: $E(X) = M_1 = \frac{1}{n} \hat{\sum}_{i=1} X_i = \bar{X}$
- 2^{nd}: $E(X^2) = M_2 = \frac{1}{n} \hat{\sum}_{i=1} X_i^2$

② solve for the parameters

③ Results = METHOD OF MOMENTS ESTIMATORS (MME): The empirical distribution converges to the probability distribution. so the corresponding moments should be about equal when $n \rightarrow \infty$

• FROM THE ORIGIN: THEORETICAL MOMENT = SAMPLE MOMENT solve for

Ex: $X_i \sim$ Bernoilli (p)

• 1st $\quad E(X_i) = p \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i$ DONE!

$\hat{p}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Ex: $X_i \sim N(\mu, \sigma^2)$

• 1st $\quad E(X_i) = \mu \quad = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ DONE!

• 2nd $\quad E(X_i^2) = \sigma^2 + \mu^2 \quad = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \Rightarrow \hat{\sigma}^2_{MM} = \frac{1}{n}\sum X_i^2 - \mu^2$

$= \frac{1}{n}\sum X_i^2 - \bar{X}^2 =$

$= \frac{1}{n}\sum (X_i - \bar{X})^2$

$Var(X) = \sigma^2 = E(X^2) - E(X)^2 = E(X^2) - \mu^2 \Rightarrow E(X^2) = \sigma^2 + \mu^2$

The method of moments estimators = MLE estimators

$\hat{\sigma}_{MM} = \hat{\sigma}$; $\hat{\mu}_{MM} = \hat{\mu}$

• ABOUT THE MEAN

① Equate 1st theoretical moment about the origin with $M_1 = E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
1st sample moment about the origin

② Equate 2nd theoretical moment about the mean with $M_2^* = E(X-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
2nd sample moment about the mean

⋮ ⋮

③ $K = 3, 4 \ldots$ $\quad M_K^* = E(X-\mu)^K = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^K$

④ Solve for parameters

⑤ Result = METHOD of MOMENTS ESTIMATORS

Ex: $X_i \sim \gamma(\alpha, \theta)$.

pdf = density function: pdf of γ dist: $f(x_i) = \frac{1}{\Gamma(\alpha)\theta^\alpha} \cdot x^{\alpha-1} \cdot e^{-x/\theta}$ for $x > 0$

likelihood function: $L(\alpha, \theta) = \left(\frac{1}{\Gamma(\alpha)\theta^\alpha}\right)^n \cdot (x_1 x_2 \cdots x_n)^{\alpha-1} \cdot e^{\frac{-1}{\theta}\cdot \sum x_i}$

$L'(\alpha, \theta)$: is hard to do bc of $\Gamma(\alpha)$ so MLE is hard.

$\hat{\alpha}_{MM} = ?$ $\quad \hat{\theta}_{MM} = ?$

	THEORETICAL	SAMPLE	SOLVE FOR PARAMETER
1^{st} (origin)	$E(X_i) = \alpha\Theta$	$= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$	$\alpha = \frac{\bar{X}}{\Theta}$
2^{nd} (mean)	$Var(X_i) = E(X_i-\mu)^2 = \alpha\Theta^2$	$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$	$Var(X_i) = \frac{\bar{X}}{\Theta}\cdot\Theta^2 = \bar{X}\Theta = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

$$\hat{\Theta}_{MM} = \frac{1}{\bar{X}}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\hat{\alpha}_{MM} = \frac{\bar{X}}{\Theta} = \frac{\bar{X}}{\frac{1}{n\bar{X}}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

ASYMPTOTIC DISTRIBUTIONS OF MAXIMUM LIKELIHOOD ESTIMATORS:

Continuous distribution $\Rightarrow$ pdf $f(x;\Theta)$.

We want to find $\hat{\Theta}$ by solving $\frac{\partial(\ln L(\Theta))}{\partial\Theta} = 0$

If the approximator $\hat{\Theta}$ is close enough to Θ, by a series of proofs:

- CLT
- Taylor series.

$$\hat{\Theta} \sim N\left(\Theta, \frac{1}{-nE\left[\frac{\partial^2 \ln f(x;\Theta)}{\partial\Theta^2}\right]}\right)$$

This can be used for continuous and discrete dist

Ex: CONTINUOUS: $X_1, X_2 \dots X_n$ random sample from exponential dist.

$X_i \sim Exp(\Theta)$. pdf: $f(x;\Theta) = \frac{1}{\Theta}e^{-\frac{x}{\Theta}}$ $0<x<\infty$ $\Theta \in \Omega\ (0<\Theta<\infty)$

$$L(\Theta) = L(\Theta; x_1, x_2 \dots x_n) = \left(\frac{1}{\Theta}e^{-\frac{x_1}{\Theta}}\right)\cdot\left(\frac{1}{\Theta}\cdot e^{-\frac{x_2}{\Theta}}\right)\cdot\left(\frac{1}{\Theta}\cdot e^{-\frac{x_3}{\Theta}}\right) = \frac{1}{\Theta^n}\cdot e^{-\frac{\sum x_i}{\Theta}} \quad 0<\Theta<\infty$$

$$\ln(L\Theta) = -n\cdot\ln(\Theta) - \frac{1}{\Theta}\cdot\sum x_i$$

$$\frac{\partial \ln(L\Theta)}{\partial\Theta} = \frac{-n}{\Theta} - (-1)\cdot\frac{1}{\Theta^2}\cdot\sum x_i = -\frac{n}{\Theta} + \frac{\sum x_i}{\Theta^2} = \frac{-\Theta n + \sum x_i}{\Theta^2} = 0$$

$$\Theta = -\frac{\sum x_i}{-n} = \frac{\sum x_i}{n} = \bar{X} \text{ is the MLE.}$$

$$\ln f(x;\Theta) = \ln\left(\frac{1}{\Theta}e^{-\frac{x}{\Theta}}\right) = (-1)\cdot\ln(\Theta) \quad \frac{-x}{\Theta}\cdot\ln(e) = -\ln(\Theta) - \frac{x}{\Theta}$$

$$\frac{\partial(\ln f(x;\Theta))}{\partial\Theta} = -\frac{1}{\Theta} - x\cdot(-1)\cdot\frac{1}{\Theta^2} = -\frac{1}{\Theta} + \frac{x}{\Theta^2}$$

$$\frac{\partial^2(\ln f(x;\Theta))}{\partial\Theta} = (-1)\cdot(-1)\cdot\frac{1}{\Theta^2} + x\cdot(-2)\cdot\frac{1}{\Theta^3} = \frac{1}{\Theta^2} - \frac{2x}{\Theta^3}$$

$$-E\left(\frac{\partial^2(\ln f(x;\Theta))}{\partial\Theta}\right) = E\left(\frac{1}{\Theta^2} - \frac{2x}{\Theta^3}\right) = -\frac{1}{\Theta^2} + \frac{2}{\Theta^3}\cdot E(x) = -\frac{1}{\Theta^2} + \frac{2}{\Theta^3}\cdot\Theta = \frac{1}{\Theta^2}$$

$$\left(E(x) = \bar{x} = \Theta\right)$$

VARIANCE and also the CRLB

$$\hat{\Theta} \sim N\left(\Theta, \frac{1}{-nE\left[\frac{\partial^2 \ln f(x;\Theta)}{\partial\Theta^2}\right]}\right) \Rightarrow \hat{\Theta} \sim N\left(\Theta, \underbrace{\frac{1}{n\cdot\frac{1}{\Theta^2}} = \frac{\Theta^2}{n}}\right)$$

mean standard deviation $= \frac{\Theta}{\sqrt{n}}$

The variance of $\bar{X}$ = the CRLB (lower bound)

↳ $\bar{X}$ is an unbiased minimum variance estimator (UMVUE)

$\hat{\Theta}$

The random interval $\bar{X} \pm 1.96\cdot\frac{\Theta}{\sqrt{n}}$ has a prob of 95% that it covers Θ.

$\bar{x} \pm 1.96\cdot\frac{\bar{x}}{\sqrt{n}}$ is an approximate 95% CI for Θ

Ex: Discrete: $X_i \sim$ Poisson (λ) → pmf: $f(x,\lambda) = \frac{\lambda^x\cdot e^{-\lambda}}{x!}$ $x = 0,1,2,\ldots$ $\lambda \in \mathbb{R}^+ (0<\lambda<\infty)$

MLE for $\hat{\lambda} = \bar{X}$

$$\ln f(x;\lambda) = x\cdot\ln(\lambda) - \lambda - \ln(x!)$$

$$\frac{\partial(\ln f(x;\lambda))}{\partial\lambda} = x\frac{1}{\lambda} - 1 \quad ; \quad \frac{\partial^2(\ln f(x;\lambda))}{\partial\lambda} = -\frac{x}{\lambda^2}$$

$$-E\left(-\frac{x}{\lambda^2}\right) = \frac{1}{\lambda^2}\cdot E(x) = \frac{1}{\lambda^2}\cdot\lambda = \frac{1}{\lambda}$$

VARIANCE and also the CRLB (lower bound)

$$\hat{\Theta} \sim N\left(\Theta, \frac{1}{-nE\left[\frac{\partial^2 \ln f(x;\Theta)}{\partial\Theta^2}\right]}\right) \Rightarrow \hat{\lambda} \sim N\left(\lambda, \underbrace{\frac{1}{n\cdot\frac{1}{\lambda}} = \frac{\lambda}{n}}\right)$$

mean standard dev $= \sqrt{\frac{\lambda}{n}} = \sqrt{\frac{\bar{x}}{n}}$

The variance of $\bar{X}$ = the CRLB (lower bound)

↳ $\bar{X}$ is an unbiased minimum variance estimator (UMVUE)

λ

* see next page.

UNIFORMLY MINIMUM VARIANCE UNBIASED ESTIMATOR (UMVUE)

The variance of $\hat{\theta}$ serves as the lower bound for the variance of any UNBIASED estimator of θ $\Rightarrow$ therefore, if we find an unbiased estimator that has a variance = to the lower bound of $\hat{\theta}$ $\Rightarrow$ best estimator or UMVUE. because we cannot find a better estimator.

so for $x_1, x_2 \dots x_n$ random sample of size n from dist $f(x;\theta)$ the estimator U^* is UMVUE if: $\begin{cases} \bullet\ U^* \text{ is an UNBIASED estimator of } \theta \\ \bullet\ \text{for any other unbiased estimator of } \theta\ U \\ \quad \text{Var}(U^*) \leq \text{Var}(U) \quad \text{for all } \theta \in \Omega \end{cases}$

CRAMER-RAO LOWER BOUND (CRLB) or INEQUALITY.

For $x_1, x_2 \dots x_n$ random sample of size n from continuous dist $f(x;\theta)$ $\theta \in \Omega = \{\theta; c < \theta < d\}$

The support of x does not depend on θ.

$y = u(X_1, X_2 \dots X_n)$ is and UNBIASED estimator of θ.

$$\text{Var}(y) \geq \frac{1}{n E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^2} = \frac{-1}{n E\left[\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2}\right]} = \frac{1}{I}$$

I = FISHER'S INFORMATION $= n E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^2 = -n E\left[\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2}\right]$

for continuous dist: $n \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(x;\theta)}{\partial \theta}\right]^2 f(x;\theta)\,dx \quad n \int_{-\infty}^{\infty} \left[\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2}\right] f(x;\theta)\,dx$

for discrete dist: $n \sum \left[\frac{\partial \ln f(x;\theta)}{\partial \theta}\right]^2 f(x;\theta) \quad n \sum \left[\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2}\right] f(x;\theta)$

EFFICIENCY of an estimator: ratio of the Rao-Cramer lower bound to the variance of any unbiased estimator.

Ex. an estimator with efficiency of 50% means that $\frac{1}{0.5} = 2$ times as many samples observations do needed to do as well in an estimator as can be done with the MVUE (that is a 100% efficient estimator).

ex: pdf $X = f(x;\theta) = \theta x^{\theta-1}$ $0<x<1$ $\theta \in \Omega = \{\theta : 0<\theta<\infty\}$

$\ln f(x;\theta) = \ln(\theta) + (\theta-1)\cdot \ln(x)$

$\frac{\partial(\ln f(x;\theta))}{\partial\theta} = \frac{1}{\theta} + \ln(x)$; $\frac{\partial^2(\ln f(x;\theta))}{\partial\theta^2} = (-1)\cdot\frac{1}{\theta^2} = -\frac{1}{\theta^2}$

$E\left(\frac{\partial^2(\ln f(x;\theta))}{\partial\theta^2}\right) = E\left(-\frac{1}{\theta^2}\right) = -\frac{1}{\theta^2}$

$\text{Var}(y) \geq \frac{-1}{n E\left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\right]} = \frac{-1}{n\cdot\left(-\frac{1}{\theta^2}\right)} = \frac{\theta^2}{n}$ is the greatest lower bound of the variance of any unbiased estimator of θ

MLE of $\theta = \hat{\theta}$

$L(\theta) = \theta x_1^{\theta-1}\cdot\theta x_2^{\theta-1}\cdots\theta x_n^{\theta-1} = n\theta\cdot\left(\prod_{i=1}^{n} x_i\right)^{\theta-1}$

$\ln(L\theta) = n\ln(\theta) + (\theta-1)\cdot\ln\left(\prod_{i=1}^{n} x_i\right)$

$\frac{\partial \ln(L\theta)}{\partial\theta} = \frac{n}{\theta} + \ln\left(\prod_{i=1}^{n} x_i\right) \equiv 0 \Rightarrow \frac{n + \ln\left(\prod_{i=1}^{n} x_i\right)\cdot\theta}{\theta} = 0$

$\hat{\theta} = \frac{-n}{\ln\left(\prod_{i=1}^{n} x_i\right)}$ is the MLE $\quad \hat{\theta} \sim N\left(\theta, \frac{\theta^2}{n}\right)$

lower bound = variance of $\hat{\theta}$

so $\hat{\theta}$ is the MVUE.

Properties of MLE (Summary) Under certain regularity conditions, the MLE

1. exists and is unique
2. is a consistent ($\approx$ asymptotically unbiased) estimator
3. is asymptotically normal
4. is asymptotically efficient.

$\hat{\theta}_{MLE} \sim N\left(\theta, -\frac{1}{n E\left[\frac{\partial^2 \ln f}{\partial\theta^2}\right]}\right)$

Problems: you need to know the pdf. ($f(x;\theta)$)

Definition (Mean Squared Error) If Y is an estimator of θ, then the bias is given by

$$\text{Bias}(Y) = E(Y) - \theta$$

and the mean squared error (MSE) of Y is given by

$$\text{MSE}(Y) = E[(Y-\theta)^2].$$

Theorem If Y is an estimator of θ, then

$$\text{MSE}(Y) = \text{Var}(Y) + [\text{Bias}(Y)]^2.$$

DELTA METHOD.

Estimator $\hat{\gamma}_n$ of a parameter $\gamma \rightarrow \hat{\gamma}_n \sim N\left(\gamma, \frac{\sigma_\gamma^2}{n}\right)$

$E(\hat{\gamma}_n) \approx \gamma \quad Var(\hat{\gamma}_n) \approx \frac{\sigma_\gamma^2}{n}$

Estimator $f(\hat{\gamma}_n)$ is an estimator of $f(\gamma)$

↳ Taylor series expansion of $f(\hat{\gamma}_n)$ about $f(\gamma)$ =>

$f(\hat{\gamma}_n) = f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma) + (\text{small remainder term})$

$E(f(\hat{\gamma}_n)) = E[f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma)] = f(\gamma) + f'(\gamma)\underbrace{E(\hat{\gamma}_n - \gamma)}_{E(\hat{\gamma}_n) - E(\gamma)} = f(\gamma)$

$Var(f(\hat{\gamma}_n)) = Var[f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma)] = f'(\gamma)^2 \cdot Var(\hat{\gamma}_n) = f'(\gamma)^2\left(\frac{\sigma_\gamma^2}{n}\right)$

$Var(a+bX) = V(a) + V(bX) + 2cov(a, bX)$

$V(f(\gamma)) + V(f'(\gamma)(\hat{\gamma}_n - \gamma)) + 2cov(a, bX)$

$V(f(\gamma)) = 0 \qquad 2cov(a, bX) = 0$

$$f(\hat{\gamma}_n) \sim N\left(f(\gamma), f(\gamma)^2\left(\frac{\sigma_\gamma^2}{n}\right)\right)$$

DELTA METHOD:

Standard error for odds / risk ratio. You have to take the log of the odds.

Approximation by Taylor series

$$\hat{\Theta} \sim N(\Theta, \frac{\sigma^2_{\hat{\Theta}}}{n})$$

$$f(\hat{\Theta}) \sim N(f(\Theta), f'(\Theta) SE_{\Theta}) = (f(\Theta), f'(\Theta)^2 Var(\Theta))$$

TAYLOR SERIES: $\frac{f^k(\Theta)(x-\Theta)^k}{k!}$ you can use it to approximate $f(x)=y$

$$f(x) \approx f(\Theta) + f'(\Theta)(x-\Theta) + \frac{f''(\Theta)(x-\Theta)^2}{2!} \ldots + \text{residual term that is even smaller}$$

$$E(y) = E[f(x)] + E[f'(\Theta)(x-\Theta)] = f(x) + f'(\Theta)\underbrace{(\Theta-\Theta)}_{0}$$

$$E(y) = f(x)$$

$$Var(x) \approx Var[f(\Theta) + f'(\Theta)(x-\Theta)] = Var[f'(\Theta)(x-\Theta)] =$$

$$f'(\Theta)^2 Var x - Var(\Theta) = f'(\Theta)^2 \cdot Var(x)$$

ANOTHER METHOD:

If $\hat{\Theta} \to \Theta$ $\frac{dx}{x}$ $\frac{f(\hat{\Theta}) - f(\Theta)}{\hat{\Theta}-\Theta} \approx f'(\hat{\Theta}) \Rightarrow \frac{f(\hat{\Theta}) - f(\Theta)}{f'(\Theta)} \approx \hat{\Theta} - \Theta \left(\frac{SE_{\Theta}}{SE_{\Theta}}\right)$

$$\Rightarrow \frac{f(\hat{\Theta}) - f(\Theta)}{f'(\Theta) SE_{\Theta}} \approx \frac{\hat{\Theta}-\Theta}{SE_{\Theta}}$$

CONFIDENCE INTERVALS FOR ONE MEAN

C.I :
- of the mean: {1 sample: $P(Y=1)$; 2 samples: z, t, bin
- of the variance: {1 sample: χ^2; 2 samples: F

Problem?

- When we use the sample {mean $\bar{x}$; proportion $\hat{p}$} to estimate the population {mean μ; proportion p}

can we be confident that $\bar{x}$ / $\hat{p}$ is close to μ / p ?

- And how close is the sample statistic to the pop. parameter?

Instead of using a point estimate, we can use an interval of values that we are confident contains the actual unknown pop. parameter.

CONFIDENCE INTERVAL: the interval of lower (L) and upper (U) values between which we can be really confident the population mean falls. $L < \mu < U$ / $L < p < U$

CONFIDENCE COEFFICIENT: reported as a proportion: $1-\alpha$: 0.9, 0.95, 0.99

CONFIDENCE LEVEL: reported as a percentage: $(1-\alpha)\cdot 100$: 90%, 95%, 99%

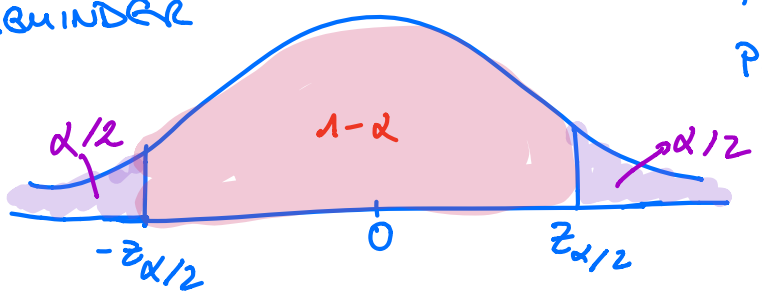
"we can be 95% confident that the population falls between L and U.

CONFIDENCE LIMITS: endpoints a and b of a CI

CRITICAL VALUES: cut-off values. ex: $z_{\alpha/2}$

COVERAGE PROBABILITY (CP): CP of the random interval [L,U] is the probability that [L,U] covers the true θ and is denoted by $P(\theta \in [L,U])$

REMINDER



$P(Z > z_{\alpha/2}) = \alpha/2$

$P(-Z < z_{\alpha/2}) = 1 - P(Z > z_{\alpha/2}) = \alpha/2$

CONFIDENCE INTERVAL FOR POPULATION MEAN.

- ASSUME σ IS KNOWN:

-Generic confidence interval for the mean: $P(a < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < b) = 1-\alpha$

$$P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$$

-Z-interval for a mean:

Assume:

- $X_i = X_1, X_2 \dots X_n \quad X_i \sim N(\mu, \sigma^2)$ and $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- The population variance σ^2 is known.

$(1-\alpha)\cdot 100\%$ confidence interval for the mean μ is $\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Because μ is the parameter of interest, we isolate μ in the middle.

$$P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = P(-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} - \bar{x} < -\mu < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} - \bar{x}) =$$

$$= P(\bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

A common choice of $\alpha = 0.05 \rightarrow z_{\alpha/2} = 1.96$.

We can be $(1-\alpha)\cdot 100 = (1-0.05)\cdot 100 = 95\%$ confident that the μ is in the interval $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

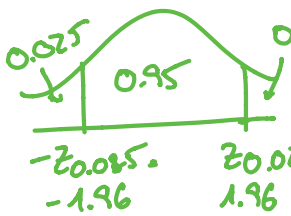
Ex: sample $n = 126$ policemen whole average 29.2 µg/dl blood lead level

X = blood lead level. $X \sim N(\mu; \sigma = 7.5 \mu g/dl)$

Historically avg blood level of women w/ no exposure = 18.2 µg/dl.

Do the policemen have elevated blood concentrations?

95% CI $\rightarrow 1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha/2 \rightarrow 0.025 \rightarrow z_{0.025} = 1.96$.



Z-interval knowing: $\bar{x} = 29.2 \mu g/dl$, $\sigma = 7.5 \mu g/dl$, $n = 126$

$$\left[29.2 - 1.96\cdot\frac{7.5}{\sqrt{126}}, 29.2 + 1.96\cdot\frac{7.5}{\sqrt{126}}\right] = [27.89, 30.51]$$

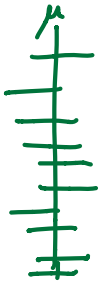
We are 95% confident that the mean blood level is between [27.89, 30.51]

Because [27.89, 30.51] does not include the historical avg 18.2 µg/dl → policemen is elevated.

INTERPRETATION:

correct: we are $(1-\alpha)\%$ confident that the interval includes the true pop. mean μ

Ex: if we took 1000 samples → calculated 95% CI for each → 950 of the 1000 intervals would actually contain the unknown value μ.



if we take only 1 sample, the interval will either contain or not the μ.

So we can only say we are 95% confident the interval contains μ.

incorrect: the probability that the pop mean μ falls between L and U is $1-\alpha$

because: the pop. μ is a constant, not a random variable.

you can't make a probability statement about a constant that doesn't change, only about random vbles.

Ex: so if you say $P[30<\mu<90]=95\%$ that's not correct. if $\begin{cases}\mu=40 \rightarrow P[30<\mu<90]=1\\ \mu=100 \rightarrow P[30<\mu<90]=0\end{cases}$

LENGTH OF AN INTERVAL: we want the interval to be as narrow as possible.

if the interval is $L<\theta<U$ → Length $U-L$.

For the z-interval: Length $\left[\bar{x}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]-\left[\bar{x}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]=2\cdot z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Length of the interval is influenced by:

- standard deviation σ: as $\sigma\downarrow$ → length ↓
- sample size n: as $n\uparrow$ → length ↓
- confidence level (z): as confidence level ↓ → length ↓

Ex.: 95% → $z=1.96$. 90% → $z=1.645$

- ASSUME σ IS UNKNOWN.

we estimate the population st.dev. σ from the sample standard deviation S.

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

For $x_1, x_2 \dots x_n \rightarrow x_i \sim N(\mu, \sigma^2)$ then a $(1-\alpha)\cdot 100\%$ confidence interval for μ is

-t- interval for the mean: $\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$

To derive the confidence interval we use $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ instead of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

For $x_1, x_2 \dots x_n \rightarrow x_i \sim N(\mu, \sigma^2)$ then $T=\frac{\bar{x}-\mu}{S/\sqrt{n}}$ follows t dist w/ $(n-1)$ df

PROOF:

a T random variable $T = \frac{Z}{\sqrt{\frac{U}{r}}}$ → $Z \sim N(0,1)$, $U \sim \chi^2_{(r)}$ } if independent $\Rightarrow T \sim t(r)$

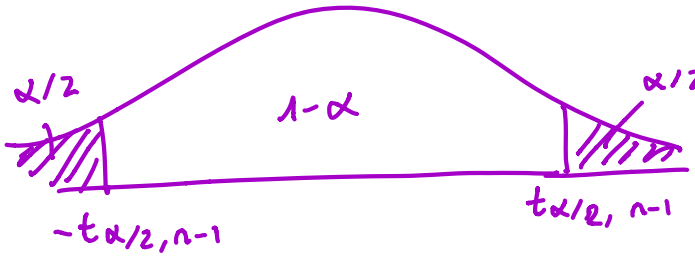
For $X_i = X_1, X_2, \dots X_n \Rightarrow X_i \sim N(\mu, \sigma^2) \Rightarrow$

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ } Z and S^2 are indep

$$T = \frac{Z}{\sqrt{\frac{U}{r}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$



$$P\left[-t_{\alpha/2,n-1} < T = \frac{\bar{X}-\mu}{S/\sqrt{n}} < +t_{\alpha/2,n-1}\right] = 1-\alpha$$

$$\bar{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}$$

- POINT ESTIMATE: $\bar{X}$ is a point estimate of μ
- INTERVAL ESTIMATE: $\bar{X} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$
  - MARGIN OF ERROR: $t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$
  - STANDARD ERROR OF THE MEAN: $\frac{S}{\sqrt{n}}$

ONE SIDED CONFIDENCE INTERVAL:

Replace $z, t_{\alpha/2}$ with z, t_{α} $\left[\bar{X} - z, t_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right)$ or $\left(-\infty, \bar{X} + z, t_{\alpha}\frac{\sigma}{\sqrt{n}}\right]$

For exercises, 1st you'd have to make sure your data is normal
2nd find your t value: ex $n=16$ 95% CI $\Rightarrow t_{0.025,15} = 2.1314$
3rd Apply formula.

NON-NORMALLY DISTRIBUTED DATA

- with sample ≥ 30 (although possibly more depending on skewness of dist.)
  $\bar{X} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$ and $\bar{X} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$ give similar results.

- with sample < 30:
  - use $\bar{X} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$ with caution or
  - use non-parametric CI

BAYESIAN INTERPRETATION of the interval estimator:

	FREQUENTIST	BAYESIAN
95% coverage	in 95% of repeated experments the intervals will cover the true Θ	the pobability that the Θ is in the interval is 95%
Randomness	repetition of experments	uncertainty about the value of the Θ

"CREDIBLE INTERVAL"

CONFIDENCE INTERVALS FOR 2 MEANS

- **KNOWN VARIANCES:**

$x_i = x_1, x_2 \dots x_n \to x_i \sim N(\mu_x, \sigma_x^2)$. $y_i = y_1, y_2 \dots y_m \to y_i\ (\mu_y, \sigma_y^2)$

σ_x^2 and σ_y^2 are known. we want to compare μ_x and μ_y

$\bar{x} - \bar{y}$ is $N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} - \frac{\sigma_y^2}{n}\right) \to P\left((\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{n}} < \mu_x - \mu_y < (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{n}}\right)$

$$P\left(-z_{\alpha/2} < \frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}}} < z_{\alpha/2}\right) = 1-\alpha$$

- **UNKNOWN BUT EQUAL VARIANCES:**

$x_i = x_1, x_2 \dots x_n \to x_i \sim N(\mu_x, \sigma_x^2)$. $y_i = y_1, y_2 \dots y_m \to y_i\ (\mu_y, \sigma_y^2)$

x_i and y_i are independent random variables.

$(1-\alpha)\,100\%$ confidence interval for $\mu_x - \mu_y$ the difference in the population means is:

$$(\bar{x}-\bar{y}) \pm t_{\alpha/2,(n+m-2)}\, s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where s_p^2 = POOLED SAMPLE VARIANCE is an unbiased estimator of the common variance σ^2.

it's an average of the sample variances weighted by the sample sizes.

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

ASSUMPTIONS:

- x_i and y_i are independent
- the measurements of each pop are normally distributed.
- the measurements of each pop have the same variance σ^2

INTERPRETATIONS:

if the difference of the sample means CI contains 0, we cannot conclude that the pop. means differ.

PROOF:

$\bar{x} \sim N\left(\mu_x, \frac{\sigma^2}{n}\right)$; $\bar{y} \sim N\left(\mu_y, \frac{\sigma^2}{m}\right) \Rightarrow \bar{x} - \bar{y} \sim N\left(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$ because x,y independent

$$Z = \frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0,1)$$

$$\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1} \; ; \; \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1}$$

bc x,y are normal $U = \frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{n+m-2}$

$$T = \frac{Z}{\sqrt{\frac{U}{n+m-2}}} = \frac{\frac{1}{\sigma}\cdot\frac{(\bar{x}-\bar{y}) - (\mu_x-\mu_y)}{\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\frac{1}{\sigma}\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{(n+m-2)}}} = \frac{\bar{x}-\bar{y} - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

s_p in the denominator $\sim t_{n+m-2}$

- UNKNOWN AND UNEQUAL VARIANCES:

- LARGE SAMPLE SIZE $(\bar{x}-\bar{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$

Proof:

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \to z \sim N(0,1)$$

Assumptions:

X and Y are normally dist or m and $n \to \infty$

- SMALL SAMPLE SIZE

WELCH'S t-interval: $\bar{x}-y \pm t_{\alpha/2,r} \cdot \sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$ $r = \frac{\left(\frac{s_x^2}{n}+\frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1}+\frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}$

Uses a modified t-dist w r df

Protects against smaller sample size associated w larger variance.

Use the integer [r]

Proof:

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \sim t(r)$$ with small sample sizes

WHEN TO POOL VARIANCE?

$\frac{s_x^2}{s_y^2} > 4$ or $\frac{s_y^2}{s_x^2} > 4$ → use welch's t-interval.

Otherwise use 2-sample t-interval.

$X \sim N(0,1)$ $Y \sim N(0,36)$ with student's t-dist superimposed

Ex:

$n=6$ $m=18$ $n<m$

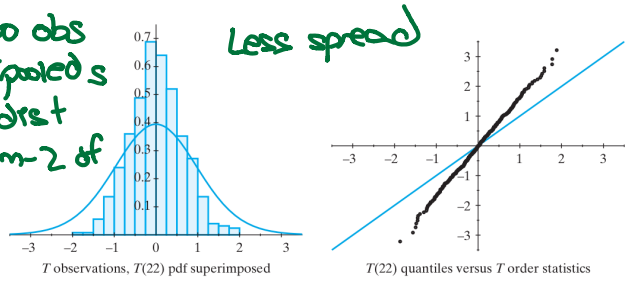
$\sigma_x^2 = 1$ $\sigma_y^2 = 36$ $\sigma_x^2 < \sigma_y^2$

$m=6$ $n=18$ $n>m$

$\sigma_x^2 = 1$ $\sigma_y^2 = 36$ $\sigma_x^2 < \sigma_y^2$

500 obs w pooled s t-dist $n+m-2$ df

Less spread



500 obs w welchs welch dist r df.

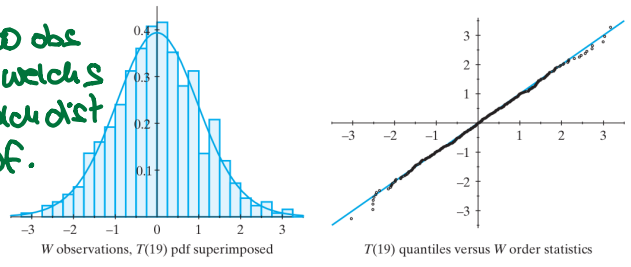
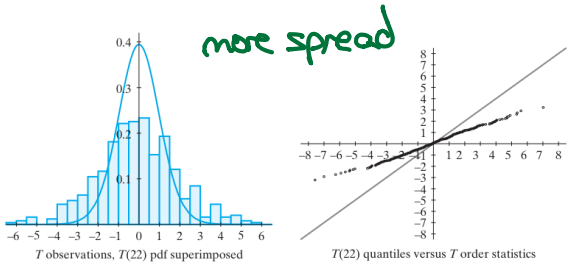


Figure 7.2-1 Observations of T and of W, $n=6$, $m=18$, $\sigma_x^2=1$, $\sigma_y^2=36$

more spread



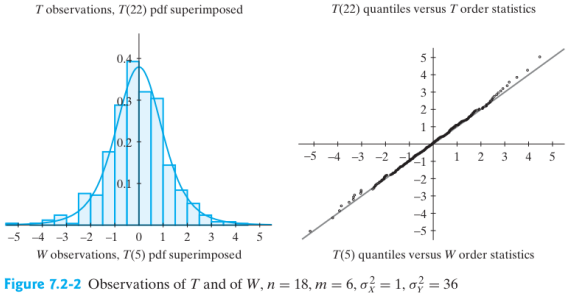


Figure 7.2-2 Observations of T and of W, $n=18$, $m=6$, $\sigma_x^2=1$, $\sigma_y^2=36$

- PAIRED SAMPLES :

we cannot use t-statistics bc x and y are not independent

$(x_1, y_1)\ (x_2, y_2)\ \dots\ (x_n, y_n)$ pairs of measurements (before, after etc)

$D_i = X_i - Y_i$ $\quad D_i \sim N(\mu_D, \sigma_D^2)$ is a random sample of pairs.

$(1-\alpha)\cdot 10\%$ C.I. of the sample mean difference to estimate the pop mean difference is

$$\bar{d} \pm t_{\alpha/2,\, n-1}\left(\frac{s_d}{\sqrt{n}}\right)$$

Subject	Red (x)	Green (y)	$d = x - y$
1	0.30	0.43	–0.13
2	0.23	0.32	–0.09
3	0.41	0.58	–0.17
4	0.53	0.46	0.07
5	0.24	0.27	–0.03
6	0.36	0.41	–0.05
7	0.38	0.38	0.00
8	0.51	0.61	–0.10

$\longrightarrow \bar{d} = -0.0625$

$s_d = 0.0765$

CONFIDENCE INTERVALS OF VARIANCES

ONE VARIANCE:

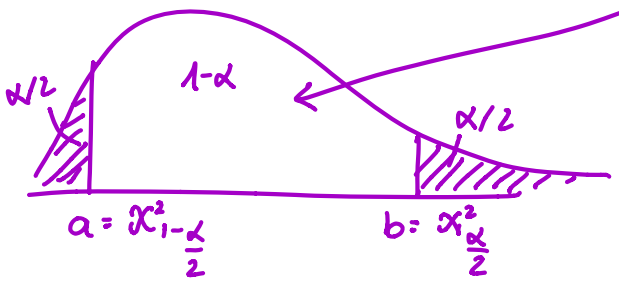
for $x_i = x_1, x_2 \dots x_n$ $\quad X_i \sim N(\mu, \sigma^2)$ with $\quad a = \chi^2_{1-\alpha/2,\, n-1} \quad b = \chi^2_{\alpha/2,\, n-1}$

$(1-\alpha)\%$ confidence interval for

- the pop variance σ^2 $\left(\dfrac{(n-1)s^2}{b} \leq \sigma^2 \leq \dfrac{(n-1)s^2}{a} \right)$
- the standard deviation σ $\left(\dfrac{\sqrt{(n-1)s}}{\sqrt{b}} \leq \sigma \leq \dfrac{\sqrt{(n-1)s}}{\sqrt{a}} \right)$

Proof

We know that if $x_i \sim N(\mu, \sigma^2)$ $\quad \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$



$$P\left[a \leq \frac{(n-1)s^2}{\sigma^2} \leq b \right] = 1-\alpha$$

$$\frac{1}{a} \geq \frac{\sigma^2}{(n-1)s^2} \geq \frac{1}{b} \Rightarrow \frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \Rightarrow$$

$$\sqrt{\frac{(n-1)s^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{a}}$$

Ex: Packs of candy with target weight 52 g. Estimate σ.
sample of candy $n = 10$. sample variance $s = 4.2$. 95% CI?

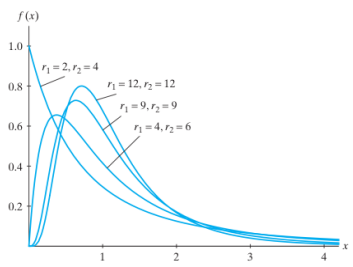
$a = \chi^2_{1-\alpha/2,\, n-1} = \chi^2_{0.975,\, 9} = 2.7 \qquad b = \chi^2_{\alpha/2,\, n-1} = \chi^2_{0.025,\, 9} = 19.02$

$$\left(\frac{\sqrt{9} \cdot 4.2}{\sqrt{19.02}} < \sigma^2 < \frac{\sqrt{9} \cdot 4.2}{\sqrt{2.7}} \right) \Rightarrow \left(1.41 < \sigma < 3.74 \right)$$

F-DISTRIBUTION:

$U \sim \chi^2(r_1) \qquad V \sim \chi^2(r_2)$. U and V independent. $\qquad F = \dfrac{\frac{U}{r_1}}{\frac{V}{r_2}} \sim F(r_1, r_2)$

$P[W \geq F_\alpha(r_1, r_2)] = \alpha.$

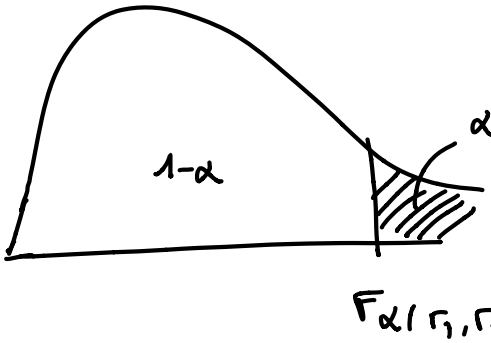


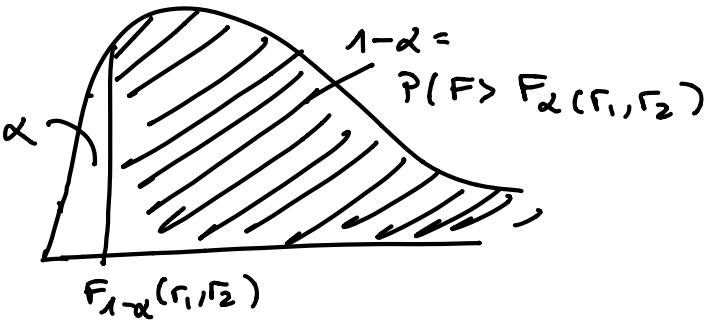
$$pdf = \frac{(r_1/r_2)^{r_1/2}\,\Gamma[(r_1+r_2)/2]\,w^{r_1/2-1}}{\Gamma(r_1/2)\,\Gamma(r_2/2)\,[1+(r_1 w/r_2)]^{(r_1+r_2)/2}},$$

if $\omega \sim F(r_1, r_2) \Rightarrow \dfrac{1}{\omega} \sim F(r_2, r_1)$

$$\alpha = P\left[\omega \leq F_{1-\alpha}(r_1, r_2)\right] = P\left[\frac{1}{\omega} \geq \frac{1}{F_{1-\alpha}(r_1, r_2)}\right]$$

$$\frac{1}{F_{1-\alpha}(r_1, r_2)} = F_\alpha(r_2, r_1) \quad \text{OR} \quad F_{1-\alpha}(r_1, r_2) = \frac{1}{F_\alpha(r_2, r_1)}$$





upper $100 \cdot \alpha^{th}$ percentile of a F dist is $F_\alpha(r_1, r_2)$

RATIO OF 2 VARIANCES :

x_i and y_i are independent, normally distributed random samples

$x_i = x_1, x_2 \dots x_n \sim N(\mu_x, \sigma_x^2)$ $\quad y_i = y_1, y_2 \dots y_m \sim N(\mu_y, \sigma_y^2)$

$c = F_{1-\alpha/2}(m-1, n-1) = \dfrac{1}{F_{\alpha/2}(n-1, m-1)}$; $d = F_{\alpha/2}(m-1, n-1)$

The $(1-\alpha)\cdot 100\%$ CI of $\dfrac{\sigma_x^2}{\sigma_y^2}$ is $\dfrac{1}{F_{\alpha/2}(n-1, m-1)} \cdot \dfrac{S_x^2}{S_y^2} \leq \dfrac{\sigma_x^2}{\sigma_y^2} < F_{\alpha/2}(m-1, n-1) \cdot \dfrac{S_x^2}{S_y^2}$

PROOF :

$x_i \sim N(\mu_x, \sigma_x^2) \Rightarrow \dfrac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1}$ $\quad y_i \sim N(\mu_y, \sigma_y^2) \Rightarrow \dfrac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1}$

$$F = \frac{\dfrac{\dfrac{(m-1)S_y^2}{\sigma_y^2}}{m-1}}{\dfrac{\dfrac{(n-1)S_x^2}{\sigma_x^2}}{n-1}} = \frac{\sigma_x^2}{\sigma_y^2} \cdot \frac{S_y^2}{S_x^2} \sim F(m-1, n-1)$$

$$P\left[F_{1-\frac{\alpha}{2}}(m-1, n-1) \leq \frac{\sigma_x^2}{\sigma_y^2} \cdot \frac{S_y^2}{S_x^2} \leq F_{\frac{\alpha}{2}}(m-1, n-1)\right] = 1-\alpha$$

$$P\left[\frac{1}{F_{\frac{\alpha}{2}}(n-1, m-1)} \cdot \frac{S_x^2}{S_y^2} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq F_{\frac{\alpha}{2}}(m-1, n-1)\right]$$

CI for variances are "sensitive" to the normality assumption, are not "robust"; are not accurate when data is not normally distributed.
However it can be important when we are testing if variances are equal, like in a t-test, or checking the assumptions of ANOVA.

REMINDER

NORMAL APPROXIMATION TO BINOMIAL

$Y \sim b(n,p)$ is the sum of n $X_i \sim$ Bernoilli (p) random variables. $X_1, X_2 \dots X_n$

$$Y = \sum_{i=1}^{n} X_i$$

	mean	variance
$X_i \sim$ Bernoilli (p) with	$\mu = EX = p$	$\sigma^2 = Var(X) = p(1-p) = pq$
$Y \sim$ binomial (n,p) with	$\mu = np$	$\sigma^2 = np(1-p) = npq$

CLT: $Y \sim N(np, npq)$ if n is sufficiently large $\begin{cases} np \geq 5 \\ nq \geq 5 \end{cases}$

$$W = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{\frac{Y}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\bar{X} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \quad n \to \infty$$

$$W = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{Y - np}{\sqrt{n} \cdot \sqrt{pq}} = \frac{Y - np}{\sqrt{npq}} \sim N(0,1) \quad n \to \infty$$

SAMPLE PROPORTION: $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the proportion of the sample that meets condition of interest.

It can be estimated by:

$$Z = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{\frac{\sum_{i=1}^{n} X_i - np}{n}}{\sqrt{\frac{np(1-p)}{n^2}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \longrightarrow N(0,1) \quad n \to \infty$$

CONFIDENCE INTERVALS FOR PROPORTIONS

ONE PROPORTION:

For large random samples → $(1-\alpha)\cdot 100\%$ confidence interval for a population proportion p is:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ONE SIDED: $\left[0,\ \hat{p} + z_{\alpha}\sqrt{\hat{p}\hat{q}/n}\right]$ or $\left[\hat{p} - z_{\alpha}\sqrt{\hat{p}\hat{q}/n},\ 1\right]$

CONDITIONS:

- np = nº of expected successes ≥ 5 AND
- $n(1-p)$ = nº of expected failures ≥ 5

Proof:

CLT $\Rightarrow$ for a large n $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows a $N(0,1)$

$\mu = E(x_i) = p$

$\sigma^2 = Var(x_i) = p(1-p)$

$\bar{x} = \hat{p}$

$\Rightarrow$

for a large n $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \Rightarrow P\left[-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right] \approx 1-\alpha$

$$P\left[-\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq -p \leq -\hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] =$$

$$P\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]$$

we need to know the population proportion p to estimate population proportion p !!

Instead of population proportion p we will use sample proportion $\hat{p}$ to get the approximate.

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ex: $n = 418$ Pennsylvanians surveyed on opinion. $4 = 280$ blamed doctors.

$\hat{p} = \frac{280}{418} = 0.67$. is the sample proportion.

Estimate proportion of all pennsylvanians who blame doctors with a 95% CI

$x_i = 1 \rightarrow$ randomly selected Pennsylvania does have doctors

$x_i = 0 \rightarrow$ randomly selected Pennsylvania does not have doctors.

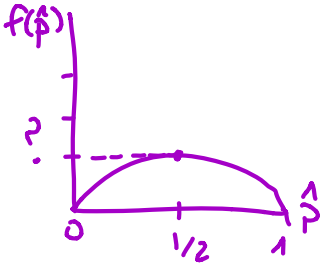
Number of P. who have docs $= \sum_{i=1}^{418} x_i = 280 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{280}{418} = 0.67.$

$z_{0.025} = 1.96 \Rightarrow 0.67 \pm 1.96\sqrt{\frac{0.67\,(1-0.67)}{418}} = 0.67 \pm 0.045$

we are 95% confident that (0.625, 0.715) Pennsylvanians have doctors

Margin of error: $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ in example: $0.045 = 4.5\%$

Maximum of $\hat{p}(1-\hat{p}) = \frac{1}{4}$ $\Rightarrow$ PROOF, $f(\hat{p})$



$f(\hat{p}) = \hat{p}(1-\hat{p})$

peak value: derivative $= 0$

$\frac{d(\hat{p}-\hat{p}^2)}{d\hat{p}} = 1 - 2\hat{p} = 0$

$\hat{p} = \frac{1}{2}$

$f(\hat{p} = \frac{1}{2}) = \frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}$

The largest margin of error is based on n: $z_{\alpha/2}\sqrt{\frac{1}{4}}\cdot\sqrt{\frac{1}{n}}$

for a 95% CI: $1.96\sqrt{\frac{1}{4}}\cdot\sqrt{\frac{1}{n}} \approx 2\sqrt{\frac{1}{4}}\cdot\sqrt{\frac{1}{n}} \approx \frac{1}{\sqrt{n}}$

$\frac{1}{\sqrt{n}}$ approximates the 95% CI margin of error.

Ex. $\frac{1}{\sqrt{418}} = 0.048$ is the approximate margin of error.

WHEN n is SMALL, 4 or $n-4$ is close to 0:

$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n+4} \Rightarrow \tilde{p} \approx \frac{y+2}{n+4} = \frac{p+2}{n+4}$

Ex: $n = 40$, $y = 8$, $1-\alpha = 0.90$

$\frac{y}{n} = \hat{p} = \frac{8}{40} = 0.20 \Rightarrow 0.20 \pm \underbrace{1.64}_{z_{\alpha/2}}\sqrt{\frac{(0.2)(0.8)}{40}}$

$\tilde{p} = \frac{8+2}{40+4} = \frac{10}{44} = 0.227 \Rightarrow 0.227 \pm 1.645\sqrt{\frac{(0.227)(0.773)}{44}}$

TWO PROPORTIONS:

For large random samples a approximate $(1-\alpha)\cdot 100\%$ confidence interval for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Proof: $\hat{p}_1 = \frac{y_1}{n_1} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$; $\hat{p}_2 = \frac{y_2}{n} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$

$$\left(\hat{p}_1 - \hat{p}_2\right) \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

$$P\left[-z_{\alpha/2} \le \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \le z_{\alpha/2}\right] \approx 1-\alpha$$

etc.— same as w/ 1 proportion — we replace p with $\hat{p}$.

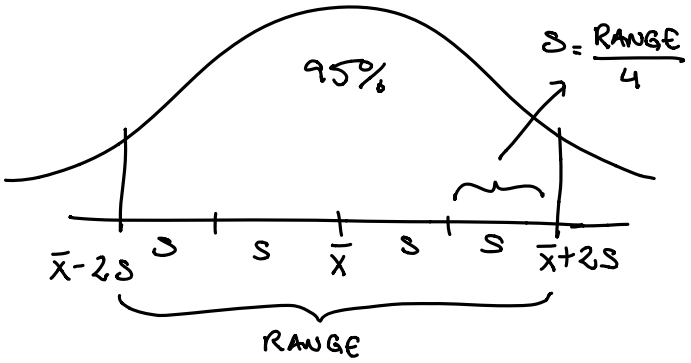
Ex:

SAMPLE SIZE

ESTIMATING A POPULATION MEAN μ (NORMAL / CONTINUOUS DATA)

Sample size necessary for estimating μ with $(1-\alpha)100\%$ CI and error $\leq \varepsilon$

$$n = \frac{(z^2_{\alpha/2})s^2}{\varepsilon^2}$$



How to determine s^2:

- scientific lit
- pilot study
- Empirical Rule: 95% of the observations to fall in interval $\bar{x} \pm 2s \Rightarrow s = \frac{MAX-MIN}{4}$
  99.7% of the observations to fall in interval $\bar{x} \pm 3s \Rightarrow s = \frac{MAX-MIN}{6}$

MAXIMUM ERROR OF THE ESTIMATE $= \varepsilon = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \Rightarrow n = \frac{z^2_{\alpha/2}\,\sigma^2}{\varepsilon^2}$

$$P\left[\bar{X} - \underbrace{z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{\varepsilon} \leq \mu \leq \bar{X} + \underbrace{z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{\varepsilon}\right] = 1-\alpha$$

Ex:

$s^2 = 10^2$	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 5$
90% ($z_{0.05} = 1.645$)	271	31	11
95% ($z_{0.025} = 1.96$)	385	43	16
99% ($z_{0.005} = 2.576$)	664	74	27

We can also change the estimate of the variance. For example, if we change the sample variance to $s^2 = 8^2$, then the necessary sample sizes for various errors ε and confidence levels $(1-\alpha)$ become:

$s^2 = 8^2$	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 5$
90% ($z_{0.05} = 1.645$)	174	20	7
95% ($z_{0.025} = 1.96$)	246	28	10
99% ($z_{0.005} = 2.576$)	425	48	17

FACTORS AFFECTING SAMPLE SIZE

- STANDARD DEVIATION (s) = $s \uparrow \rightarrow n \uparrow$ bc it is in the numerator.
- ERROR (ε) = $\varepsilon \downarrow \rightarrow n \uparrow$ bc it is in the denominator
- CONFIDENCE INTERVAL CI = CI $\uparrow \rightarrow n \uparrow$ bc it $z_{\alpha/2}$ that is in the numerator

Ex: estimate μ the mean systolic blood pressure of adult Americans w 95% CI
ε no larger than 3mm Hg. How many adults should be sampled = n ?
say sample variance $s^2 = 10^2$

$\bar{x} \pm \varepsilon = \bar{x} \pm 3$

CI for a pop. mean $\mu = \bar{x} + t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \Rightarrow \varepsilon = t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \Rightarrow$ solve for n

$$n = \frac{(t_{\alpha/2,n-1})^2 \cdot s^2}{\varepsilon^2} \Rightarrow n \approx \frac{(z^2_{\alpha/2})s^2}{\varepsilon^2} = \frac{(1.96)^2(10)^2}{3^2} = 42.7 \approx 43 \text{ people}$$

As $n \uparrow \rightarrow$ t distribution approaches the $N(0,1)$ dist.
can't use t value dependent on n if we're trying to discover n.

ESTIMATING A PROPORTION (BINARY DATA)

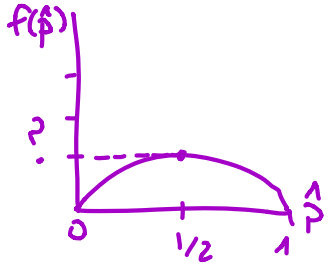
- LARGE POPULATION

SAMPLE SIZE needed to estimate a pop. proportion p with $\begin{cases}(1-\alpha)\cdot 100\% \text{ CI} \\ \text{error} \le \varepsilon\end{cases}$

$$n = \frac{z_{\alpha/2}^2\,\hat{p}(1-\hat{p})}{\varepsilon^2} \qquad \varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

How to determine $\hat{p}(1-\hat{p})$

- use prior knowledge
- set $\hat{p} = \frac{1}{2} = 0.5 \Rightarrow p(1-\hat{p}) = \frac{1}{4}$



$f(\hat{p}) = \hat{p}(1-\hat{p})$

peak value: derivative = 0

$$\frac{d(\hat{p}-\hat{p}^2)}{d\hat{p}} = 1-2\hat{p} = 0$$

$$\hat{p} = \frac{1}{2}$$

$$f(\hat{p}=\tfrac{1}{2}) = \tfrac{1}{2}(1-\tfrac{1}{2}) = \tfrac{1}{4}$$

Ex: Pollster wants to estimate p = the proportion of Americans who are Dems.

95% CI $\quad \varepsilon \le 0.03. \quad n = ?$

$$\hat{p} \pm \varepsilon = \hat{p} \pm 0.03$$

$$n = \frac{(1.96)^2\cdot(0.8)(0.2)}{0.03^2}$$

$$n = \frac{(1.96)^2(0.5)(0.5)}{0.03^2}$$

$\hat{p} = 0.8$	$\varepsilon = 0.01$	$\varepsilon = 0.03$	$\varepsilon = 0.05$
90% ($z_{0.05} = 1.645$)	4330	482	174
95% ($z_{0.025} = 1.96$)	6147	683	246
99% ($z_{0.005} = 2.576$)	10618	1180	425

$\hat{p} = 0.5$	$\varepsilon = 0.01$	$\varepsilon = 0.03$	$\varepsilon = 0.05$
90% ($z_{0.05} = 1.645$)	6766	752	271
95% ($z_{0.025} = 1.96$)	9604	1068	385
99% ($z_{0.005} = 2.576$)	16590	1844	664

- SMALL POPULATION:

$$n = \frac{m}{1+\frac{m-1}{N}} \qquad m = \frac{z_{\alpha/2}^2\,\hat{p}(1-\hat{p})}{\varepsilon^2} = \text{sample size needed for large pop.}$$

Approximate $(1-\alpha)\cdot 100\%$ CI for small population proportion is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\cdot\frac{N-n}{N-1}}$$

Proof: we take random sample $x_i = x_1, x_2 \cdots x_n$ of size n from a pop of size N.

Ex. $N = 2000$. N_1 = yes $\quad N-N_1$ = NO $\Rightarrow p = \frac{N_1}{N} \quad 1-p = \frac{N-N_1}{N}$

X = no of respondents $\quad x_i = 1$ = yes $\quad x_i = 0$ = NO

who say yes = $\sum_{i=1}^{n} x_i$

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \hat{p} = \frac{X}{n}$$

$E(x) = n \cdot \frac{N_1}{N} = np$ $\quad Var(x) = np(1-p)\left(\frac{N-n}{N-1}\right)$

$E\hat{p} = p$ $\quad Var(\hat{p}) = \frac{p(1-p)}{n} \cdot \left(\frac{N-n}{N-1}\right)$

If $n <<<< N$

$\frac{N-n}{N-1} \approx 1$

CLT $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \sim N(0,1) \Rightarrow \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}} \Downarrow = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

a small pop. $\qquad$ a large pop

As $N \downarrow \rightarrow n \downarrow$ As the pop size ↓ so does the size of the sample.

SIMPLE LINEAR REGRESSION

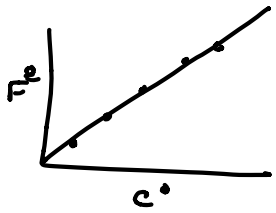
Evaluates the relationship between 2 continuous vbles:

X = independent / predictor / explanatory vble. weight → weight; speed → gas mileage; T° → evaporation

Y = dependent / response / outcome vble

DETERMINISTIC RELATIONSHIP: exact relationship $y = \alpha + \beta x$

$F^\circ = \frac{9}{5} C^\circ + 32$



STATISTICAL RELATIONSHIP: not exact relationship: $y = \alpha + \beta x + \varepsilon$

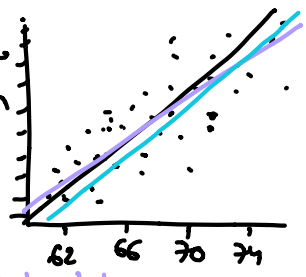
Deterministic + Random

"Trend" between x and y, but there is also some "scatter"

$$y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$$

$$y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$$

which is the best line?



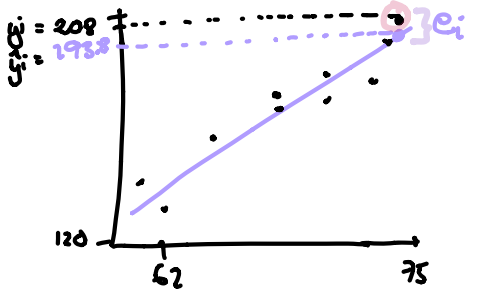
weight = -266.5 + 6.1 · height

weight = -331.2 + 7.1 x height

y_i = OBSERVED RESPONSE of i^{th} experimental unit

x_i = PREDICTOR VALUE of i^{th} experimental unit

$\hat{y}_i$ = PREDICTED RESPONSE of i^{th} experimental unit
FITTED or VALUE



$x_i = 75$

$y_i = 208$

$\hat{y}_i = -266.5 + 6.1 \cdot 75 = 193.8$

$e_i = y_i - \hat{y}_i$ = RESIDUAL ERROR of i^{th} experimental unit
or
PREDICTED ERROR

$\Rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

LEAST SQUARES CRITERION: you want to find a line that fits the data where e_i is as small as possible for all the observations i: $\hat{y}_i = a_1 + bx_i$ find a and b!

LEAST SQUARES REGRESSION LINE:

minimize the sum of the squared predicted errors

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a_1 + bx_i))^2$$

LEAST SQUARES REGRESSION LINE: $\hat{y}_i = a_1 + bx_i = a + b(x_i - \bar{x})$

$\hat{y}_i = a_1$ when $x_i = 0$ NOT CENTERED?

$\hat{y}_i = a$ when $x_i = \bar{x}$ CENTERED AROUND MEAN?

$x_i = 0$ can be not meaningful. ex: x=weight y=weight, x=0 inches ... doesn't work

LEAST SQUARES ESTIMATES

$a = \bar{y}$

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

does not change with centering

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	64	121	123.2	-2.2	4.84
2	73	181	187.1	-6.1	37.21
3	71	156	172.9	-16.9	285.61
4	69	162	158.7	3.3	10.89
5	66	142	137.4	4.6	21.16
6	69	157	158.7	-1.7	2.89
7	75	208	201.3	6.7	44.89
8	71	169	172.9	-3.9	15.21
9	63	127	116.1	10.9	118.81
10	72	165	180.0	-15.0	225.00
					766.51

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	64	121	126.271	-5.3	28.09
2	73	181	181.509	-0.5	0.25
3	71	156	169.234	-13.2	174.24
4	69	162	156.959	5.0	25.00
5	66	142	138.546	3.5	12.25
6	69	157	156.959	0.0	0.00
7	75	208	193.784	14.2	201.64
8	71	169	169.234	-0.2	0.04
9	63	127	120.133	6.9	47.61
10	72	165	175.371	-10.4	108.16
					597.29

This line would be better. But there are a 1000 lines!

PROOF:

Minimize $Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a + b(x_i - \bar{x})))^2$

$\frac{\partial Q}{\partial a} = \frac{2 \sum_{i=1}^{n} (y_i - (a + b(x_i - \bar{x}))) \cdot (-1)}{2} \overset{\text{SET}}{=} 0$ $\Rightarrow$ solve for a

$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a + b \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{0} = -\sum_{i=1}^{n} y_i + na = 0 \Rightarrow a = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$

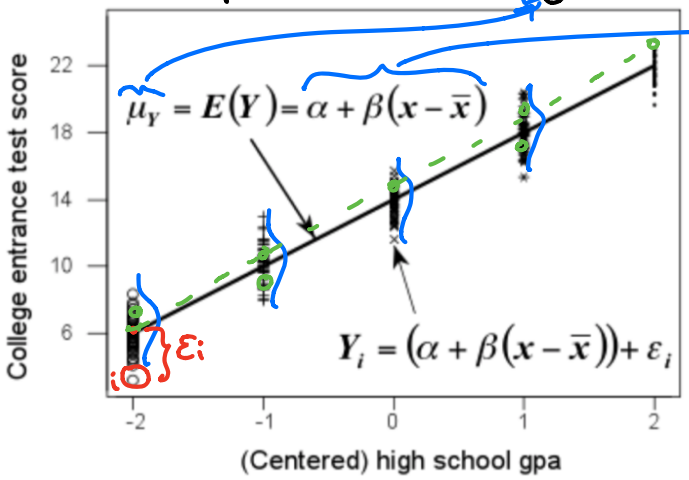
Replace a for $\bar{y}$ in Q formula: $Q = \sum_{i=1}^{n} (y_i - (\bar{y} + b(x_i - \bar{x})))^2$

$\frac{\partial Q}{\partial b} = \cancel{2} \cdot \sum_{i=1}^{n} (y_i - (\bar{y} + b(x_i - \bar{x}))) \cdot -(x_i - \bar{x}) \overset{\text{SET}}{=} 0$

$\sum_{i=1}^{n} (y_i - \bar{y} - b(x_i - \bar{x})) \cdot -(x_i - \bar{x}) = -\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + b \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$

$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Linear relationship between college entrance score and centered HS GPA ω {intercept α, slope β



Individual students i w the same centered HS GPA deviate from the mean college entrance score by ε_i

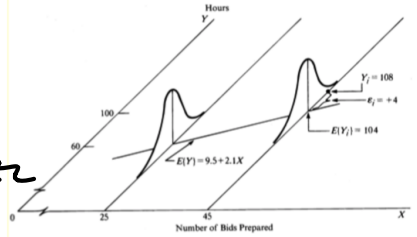
$\{ y_i = \alpha + \beta(x - \bar{x}) + \varepsilon_i$

we use a sample of students to estimate the α and β and the best fitting line.

$\{ \hat{y}_i = a + b(x_i - \bar{x})$

x = high school GPA

$x - \bar{x}$ = centered high school GPA
($\pm$ 2 points from the mean)



ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION model

- The mean of the responses $E(y_i)$ is a Linear function of x_i
- The errors ε_i, and the responses y_i are Independent
- The errors ε_i, and the responses y_i are Normally dist. => Residual plot Assess model fit
- The errors ε_i, and the responses y_i have Equal variances for all x values

$y_i - \hat{y}_i$ vs x_i

LIVE!

MAXIMUM LIKELIHOOD ESTIMATORS OF THE SIMPLE LINEAR REGRESSION PARAMETERS

if the 4 assumptions hold true:

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$
CENTERED	$a = \bar{y}$	$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n} = \frac{\sum\left[y_i-\hat{\alpha}-\hat{\beta}(x_i-\bar{x})\right]^2}{n}$ $= \frac{SS_E}{n}$
NOT CENTERED	$a = \bar{y} - \hat{\beta}\bar{x}$	SAME	$\hat{\sigma}^2 = \frac{\sum\left[y_i-\hat{\alpha}-\hat{\beta}(x_i-\bar{x})\right]^2}{n}$

Proof:

Simple linear regression model states that $\varepsilon_i \xrightarrow{\text{independent}} \sim N(0,1)$

$y_i = \alpha + \beta(x_i-\bar{x}) + \varepsilon_i \Rightarrow y_i \sim N(\alpha+\beta(x_i-\bar{x}), \sigma^2)$.

$x_1, x_2 \dots x_n$ random sample from normal dist $x_i \sim N(\mu, \sigma^2)$

MLE of μ and σ^2 ?

pdf of normal dist $\left.\begin{array}{l} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{array}\right\}$ $f(x_i;\theta_1,\theta_2) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} \cdot e^{\left[-\frac{(x_i-\theta_1)^2}{2\theta_2}\right]}$ $\quad -\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$

LIKELIHOOD FUNCTION: $L(\theta_1,\theta_2) = \prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) = \theta_2^{-n/2}\cdot(2\pi)^{-n/2}\cdot e^{\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right]}$

$\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta(x_i-\bar{x}))^2\right]$

Likelihood fucks: $L_{y_i}(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} L_{y_i}(\alpha,\beta,\sigma^2) = (\sigma^2)^{-n/2}\cdot(2\pi)^{-n/2}\cdot e$

$\ln(L_{y_i}) = -\frac{n}{2}\sigma^2 - \frac{n}{2}\cdot 2\pi \ominus \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta(x_i-\bar{x}))^2$

to maximize $\ln(L_{y_i})$, we need to minimize $\sum_{i=1}^{n}(y_i-\alpha-\beta(x_i-\bar{x}))^2 \Rightarrow$

with respect to α and β with respect to α and β

$\frac{\partial \ln(L_{y_i})}{\partial \alpha} = 0 \Rightarrow a$

$\frac{\partial \ln(L_{y_i})}{\partial \beta} = 0 \Rightarrow b$

LEAST SQUARES CRITERION!!

So the Maximum Likelihood Estimators (MLE) of α and β are the least squares estimators (LSE) of α and β under the simple linear regression assumptions

MAXIMUM LIKELIHOOD ESTIMATOR OF THE VARIANCE σ^2

UNBIASED

- ONE POPULATION

$S^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$

- MANY POPULATIONS

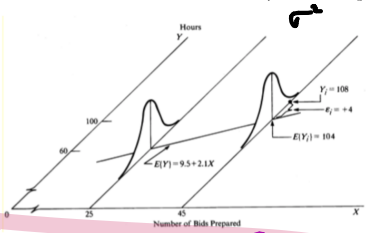
MEAN SQUARE ESTIMATOR

MSE: $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$

BIASED.

MAX LIKELIHOOD ESTIMATOR

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n}$



Proof:

$\frac{\partial \ln(L_{y_i})}{\partial \sigma^2} \overset{\text{set}}{=} 0 = \left[-\frac{n}{2\sigma^2} - \frac{1}{2}\sum(y_i-\alpha-\beta(x_i-\bar{x}))^2 \cdot \frac{-1}{(\sigma^2)^2} = 0\right] \cdot 2\sigma^4$

$\Rightarrow -n\sigma^2 + \sum(y_i-\alpha-\beta(x_i-\bar{x}))^2 = 0$

q.e.d.

BLUE : Best Linear Unbiased Estimator :
estimator with the smallest variance among all the unbiased estimators

If $\begin{cases} E(y_i) = \alpha + \beta x_i \\ Var(y_i) = \sigma^2 \\ cov(y_i, y_j) = 0 \text{ for } i \neq j \end{cases}$ then the LSE $\begin{cases} \hat{\alpha} \\ \hat{\beta} \end{cases}$ are the BLUE

$\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators with finite variances

$E(\hat{\alpha}) = \alpha \qquad Var(\hat{\alpha}) = \frac{\sigma^2}{n}$

$E(\hat{\beta}) = \beta \qquad Var(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$

$y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$ → μ_y and σ_y^2 exist and are finite
→ y's are random but x's are fixed
↓ $\hat{\alpha}, \hat{\beta}$ are linear functions of $y_1, y_2 \dots y_n$

SAMPLING DISTRIBUTIONS OF THE ESTIMATORS

	$\hat{\alpha}$	$\hat{\beta}$		
CENTERED	$\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$	$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$	$cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$	$\frac{n\sigma^2}{\sigma^2} \sim \chi^2_{(n-2)}$
NOT CENTERED	$\hat{\alpha} \sim N(\alpha, \frac{\sigma^2 \sum x_i^2}{n S_{xx}})$	$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$	$cov(\hat{\alpha}, \hat{\beta}) = 0$	$\frac{n\sigma^2}{\sigma^2} \sim \chi^2_{(n-2)}$

CONFIDENCE INTERVALS FOR REGRESSION PARAMETERS

Under the assumptions of the linear regression model (LINE):

- $\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n}\right)$

Proof: Maximum Likelihood Estimator of $\alpha = \hat{\alpha} = \bar{Y}$
Least Squares estimator

$Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$

$$E(\hat{\alpha}) = E(\bar{Y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} E(\alpha + \beta(x_i - \bar{x})) = \frac{1}{n}\left[n\alpha + \beta \cdot \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{0}\right] = \frac{1}{n} \cdot n\alpha = \alpha$$

$$Var(\hat{\alpha}) = Var(\bar{Y}) = \frac{\sigma^2}{n}$$

- $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$

Proof: MLE of $\beta = \hat{\beta} = b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
LSE

$Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$

$$E(\hat{\beta}) = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sum_{i=1}^{n} E[(x_i - \bar{x})Y_i] = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot \overbrace{(\alpha + \beta(x_i - \bar{x}))}^{E(Y_i)} =$$

$$\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \left[\alpha \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{0} + \beta \sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \beta$$

$$Var(\hat{\beta}) = \left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \overbrace{(Var(Y_i))}^{\sigma^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

- $n \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$
- $a = \hat{\alpha}$; $b = \hat{\beta}$; $\hat{\sigma}^2$ => mutually independent.

- a $(1-\alpha)\cdot 100\%$ confidence interval for the slope parameter β is

$$b \pm t_{\alpha/2, n-2}\left(\frac{\sqrt{n}\,\hat{\sigma}}{\sqrt{n-2} \cdot \sqrt{\sum(x_i - \bar{x})^2}}\right) \simeq \hat{\beta} \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}$$

PROOF: $T = \frac{Z}{\sqrt{\frac{U}{r}}}$ $\rightarrow Z \sim N(0,1)$ $\rightarrow U \sim \chi^2_{(r)}$ } independent

$$\frac{\hat{\beta} - \beta}{\frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}} \sim N(0,1)$$

$\cdot \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$

} independent

$T \sim t_{(r)}$

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

$$T = \frac{\frac{\hat{\beta} - \beta}{\frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}}}{\sqrt{\frac{\frac{n\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta} - \beta \cdot \sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}} \cdot \sqrt{\frac{1}{n-2}}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{n \cdot \hat{\sigma}^2}{(n-2)\sum (x_i - \bar{x})^2}}} =$$

$$= \frac{\hat{\beta} - \beta}{\sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}} \sim t_{(n-2)}$$

- a $(1-\alpha) \cdot 100\%$ confidence interval for the slope parameter α is

$$\alpha \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{\hat{\sigma}^2}{n-2}} \Rightarrow \alpha \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{MSE}{n}}$$

$$T = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\hat{\sigma}^2}{n-2}}} \sim t_{(n-2)}$$

CORRELATION $\neq$ CAUSATION

$\neq$ REGRESSION: in correlation x, y are interchangeable

MORE REGRESSION:

ESTIMATION: To know the value of the mean response $E(Y)=\mu_Y$ for a given value x, calculate the CONFIDENCE INTERVAL for THE MEAN $E(Y)=\mu_Y$

PREDICTION: To know the value of a new observation Y_{n+1} for a given value x calculate the PREDICTION INTERVAL for THE NEW OBSERVATION Y_{n+1}

CONFIDENCE INTERVAL for THE MEAN $E(Y)=\mu_Y$

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

PROOF: point estimate for μ_Y is $\hat{y}=\hat{\alpha}+\hat{\beta}(x-\bar{x})$ and $\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$, $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$ — independent

$$\hat{Y}=E(\hat{y})=E[\hat{\alpha}+\hat{\beta}(x-\bar{x})]=E(\hat{\alpha})+(x-\bar{x})E(\hat{\beta})=$$
$$=\alpha+\beta(x-\bar{x})=\mu_Y$$

$$Var(\hat{y})=Var[\hat{\alpha}+\hat{\beta}(x-\bar{x})]=Var(\hat{\alpha})+(x-\bar{x})^2E(\hat{\beta})=\frac{\sigma^2}{n}+(x-\bar{x})^2\cdot\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=$$
$$=\sigma^2\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x-\bar{x})^2}\right]$$

$$\hat{Y} \sim N\left(\mu_Y, \sigma^2\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x-\bar{x})^2}\right]\right)$$

$$T=\frac{\frac{\hat{Y}-\mu_Y}{\sigma\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x-\bar{x})^2}}}}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}/(n-2)}}=\frac{\hat{Y}-\mu_Y}{\sqrt{MSE}\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x-\bar{x})^2}}} \sim t_{(n-2)}$$

$$MSE=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$$

$$P\left[-t_{\alpha/2,(n-2)} \le \frac{\hat{Y}-\mu_Y}{\sqrt{MSE}\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x-\bar{x})^2}}} \le t_{\alpha/2,(n-2)}\right]=1-\alpha$$

How to make the CI for the mean narrower?

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}}$$

① estimate μ_y at $\bar{x}$ = the mean of the predictor values
shortest C.I for μ_y when $\bar{x}=x \Rightarrow (x-\bar{x})$ is smallest

② ↓ confidence level → ↓ $t_{\alpha/2,n-2}$

③ ↑ sample size → ↑ n → ↓ $\frac{1}{n}$

④ choose spread out predictor values x_i → ↑ spread → ↑ $\Sigma(x_i-\bar{x})^2$

Ex: x = duration of previous eruption y = time until next eruption
mean time till next eruption = $\bar{y}$ = 8.5 min. → 95% confidence that the mean time until next eruption will be between (70.14, 72.203) min

PREDICTION INTERVAL FOR THE NEW OBSERVATION y_{n+1}

A $(1-\alpha)\cdot 100\%$ prediction interval for a new observation y_{n+1} when the predictor $x = x_{n+1}$

$$\hat{y}_{n+1} \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}}$$

PROOF: $y_{n+1} \sim N(\alpha+\beta(x_{n+1}-\bar{x}), \sigma^2)$ and

$$w = y_{n+1}-\hat{y}_{n+1} = y_{n+1}-\hat{\alpha}-\hat{\beta}(x_{n+1}-\bar{x})$$

$\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$
$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$
$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{(n-2)}$
} independent

$$E(w) = E[y_{n+1}-\hat{\alpha}-\hat{\beta}(x_{n+1}-\bar{x})] = E(y_{n+1}) - E(\hat{\alpha}) - (x_{n+1}-\bar{x})E(\hat{\beta}) =$$
$$= \alpha+\beta(x_{n+1}-\bar{x}) - \alpha - (x_{n+1}-\bar{x})\beta = 0$$

$$Var(w) = Var[y_{n+1}-\hat{\alpha}-\hat{\beta}(x_{n+1}-\bar{x})] = Var(y_{n+1}) - Var(\hat{\alpha}) - (x_{n+1}-\bar{x})^2 Var(\hat{\beta}) =$$
$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x_{n+1}-\bar{x})^2\sigma^2}{\Sigma(x_i-\bar{x})^2} = \sigma^2\left[1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}\right]$$

$$w \sim N\left(0, \sigma^2\left[1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}\right]\right)$$

$$T = \frac{\dfrac{(Y_{n+1} - \hat{Y}_{n+1}) - 0}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}}}{\sqrt{\dfrac{n\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{(Y_{n+1} - \hat{Y}_{n+1})}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$$

$$P\left[-t_{\alpha/2, n-2} \leq \frac{(Y_{n+1} - \hat{Y}_{n+1})}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \leq t_{\alpha/2, n-2}\right] = 1 - \alpha$$

Ex: x = duration of previous eruption y = time until next eruption
previous eruption = y = 3.5 min. → 95% confidence that the time until next eruption will be between (58.109, 84.784) min

- For a given value x, the prediction interval is always longer than the estimation interval, because the pred. interval has an extra term in the SE.

CONFIDENCE INTERVAL μ_y	PREDICTION INTERVAL $\hat{y}_{n+1}$
$\hat{y} \pm t_{\alpha/2, n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$	$\hat{y}_{n+1} \pm t_{\alpha/2, n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

- we can't make the standard error of the prediction for y_{n+1} approach 0

- Make narrower → same as with C.I for μ_y.

CONFIDENCE BAND : collection of all $(1-\alpha)\cdot 100\%$ C.I.'s for $\mu(x)$: $-\infty < x < \infty$
PREDICTION BAND : collection of all $(1-\alpha)\cdot 100\%$ P.I.'s for $Y(x) = \alpha + \beta x + \varepsilon$: $-\infty < x < \infty$

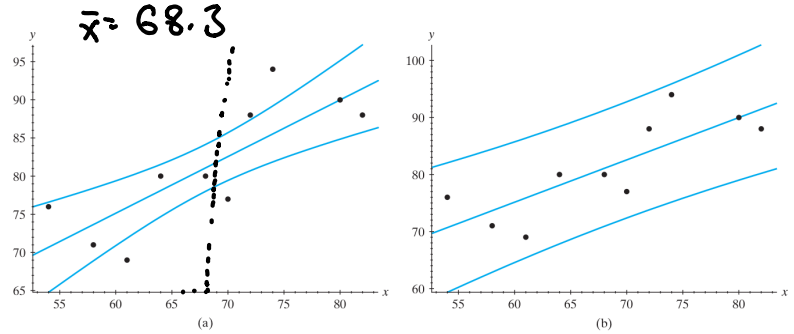


Figure 7.6-1 A pointwise 95% (a) confidence band for $\mu(x)$ and (b) prediction band for Y

The bands are narrowest at $x = \bar{x}$
Prediction band is always wider: the difference between 1 observation of y and its predictor varies more than the difference between the mean of the pop of y values and its estimator.

MULTIPLE REGRESSION

GENERAL LINEAR MODEL: $E(Y) = E(Y \mid x_1 \dots x_K) = E(Y \mid x) = \beta_1 x_1 + \beta_2 x_2 \dots \beta_K x_K$

Ex: x_1 = student's ACT score x_2 = student's HS class rank y = 1st year GPA college

To fit the linear model, $\beta_1 x_1 + \beta_2 x_2 \dots \beta_K x_K$. n observation points: $x_{1j}, x_{2j} \dots x_{Kj}$ y_j

① Minimize the function by least squares: $G = \sum_{j=1}^{n} (y_j - \beta_1 x_{1j} - \beta_2 x_{2j} - \dots \beta_K x_{Kj})^2$

② K 1st partial derivatives and $\overset{SET}{\equiv} 0$.

$$\frac{\partial G}{\partial \beta_i} = \sum_{j=1}^{n} (-2)(y_j - \beta_1 x_{1j} - \beta_2 x_{2j} \dots \beta_K x_{Kj})(x_{ij}) \overset{SET}{\equiv} 0$$

K normal equations .

$$\begin{aligned}
\beta_1 \sum_{j=1}^{n} x_{1j}^2 &+ \beta_2 \sum_{j=1}^{n} x_{1j}x_{2j} + \cdots + \beta_k \sum_{j=1}^{n} x_{1j}x_{kj} = \sum_{j=1}^{n} x_{1j}y_j, \\
\beta_1 \sum_{j=1}^{n} x_{2j}x_{1j} &+ \beta_2 \sum_{j=1}^{n} x_{2j}^2 + \cdots + \beta_k \sum_{j=1}^{n} x_{2j}x_{kj} = \sum_{j=1}^{n} x_{2j}y_j, \\
&\vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\quad \vdots \qquad\quad \vdots \\
\beta_1 \sum_{j=1}^{n} x_{kj}x_{1j} &+ \beta_2 \sum_{j=1}^{n} x_{kj}x_{2j} + \cdots + \beta_k \sum_{j=1}^{n} x_{kj}^2 = \sum_{j=1}^{n} x_{kj}y_j.
\end{aligned}$$

③ Find solutions for $\beta_1 \dots \beta_K$ => least squares estimates = MLE provided the random vbles $Y_1, Y_2 \dots Y_n$ are mutually independent and Y_j is $N(\beta_1 x_{1j} + \beta_2 x_{2j} \dots \beta_K x_{Kj}, \sigma^2)$ $j = 1 \dots n$

INTERPRETATION OF β: change in $E(Y)$ when x_1 is increased by 1 unit, while each of the $x_2 \dots x_K$ are held constant (=adjusting / controlling other x).

Caution! when data are transformed ex $\log x$, $\log y$.

HYPOTHESIS TESTING.

HYPOTHESIS: proposed explanation for a phenomenon

NULL HYPOTHESIS: H_0: initial assumption about the population parameter

ALTERNATIVE HYP. = H_1:

SIMPLE HYP.: each hypothesis has a single value

Ex: we assume $X \sim N(\mu, 36)$. $\mu = 50$ or $\mu = 55$. $H_0: \mu = 50$ then $H_1: \mu = 55$

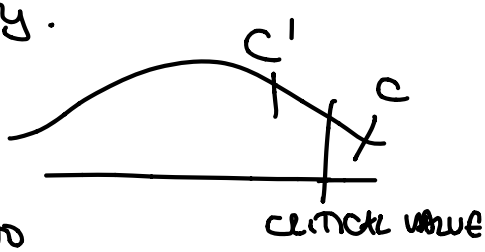
COMPOSITE HYP.: A hypothesis can be a composite of different values/distributions.

Ex: we assume $X \sim N(\mu, 36)$. $H_0: \mu = 50$; $H_1: \mu > 50 \Rightarrow \mu = 51, \mu = 52 \dots$

TEST STATISTIC: function of the sample, numerical summary.

CRITICAL REGION: Rejection region for H_0. = C

ACCEPTANCE REGION: Acceptance region for H_0. = C'



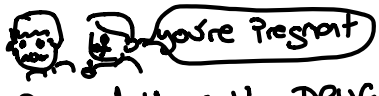
CRITICAL VALUE: threshold / cut-off value, corresponds to

SIGNIFICANCE LEVEL: probability of a Type-I error.

p-VALUE VS. CI: 95% CI ≈ interval containing all values where $p > 0.05$.

if $p < 0.05$ or CI contains 0 → reject H_0.

ERRORS



- TYPE I ERROR: α : FALSE ⊖ : Reject H_0; H_0 TRUE

  $P[(x_1, x_2, \dots x_n) \in C ; H_0] = c/a+c$

  = SIGNIFICANCE LEVEL.

- TYPE II ERROR: β : FALSE ⊕ : Accept H_0; H_0 FALSE

  $P[(x_1, x_2, \dots x_n) \in C' ; H_1] = b/b+d$

- POSTERIOR PROB. TYPE I ERROR: $c/c+d$ } BAYESIAN
- POSTERIOR PROB. TYPE II ERROR: $b/a+b$ }
- POWER FUNCTION: $1-\beta$: Reject H_0; H_1 TRUE

Table 2: True state vs. Decision

		True state	
		H_0 is true	H_0 is false
Decision	Accept H_0	a	b
	Reject H_0	c	d

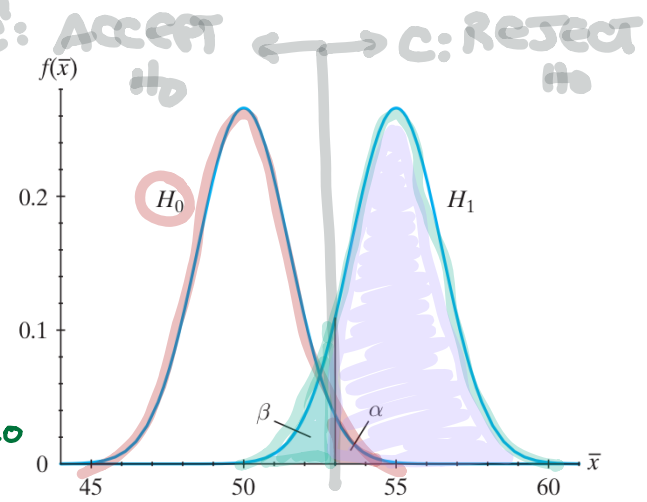


Figure 8.1-1 pdf of $\bar{X}$ under H_0 and H_1

Ex: Nine tags placed right-side. $H_0: p = 0.5$ $H_1: p < 1/2$ $n = 20$

$y = \sum_{i=1}^{20} x_i \sim bin(20, p)$ CRITICAL REGION: $C = [y \le 6] = \sum_{i=1}^{20} x_i \le 6$

$\alpha = P(y \le 6; p = \frac{1}{2}) = \sum_{y=0}^{6} \binom{20}{y} (\frac{1}{2})^{20} = 0.0577$ | $\beta = P(7 \le y \le 20; p = \frac{1}{4}) = 0.0024$ (example)

$K_p = P[y \le 6; p < \frac{1}{2}] = \sum_{y=0}^{6} \binom{20}{y} (p)^y \cdot (1-p)^{20-y}$

↑ in this example p can be any value $0 < p < 1/2$ and the power changes accordingly.

Power is a function because p can be any value and the power changes depending on the value of p.

• HYPOTHESIS TESTING: = CRITICAL VALUE APPROACH

ex: 4-sided die tossed 1000 times
290 # 4 observed. → is this more than expected?

① STATE $\begin{cases} H_0 \\ H_1 \end{cases}$

H_0: unbiased die ⇒ each side = likely ⇒ $\frac{1}{4} = 0.25 = p$.
H_1: biased die ⇒ sides ≠ likelihood ⇒ $p > 0.25$

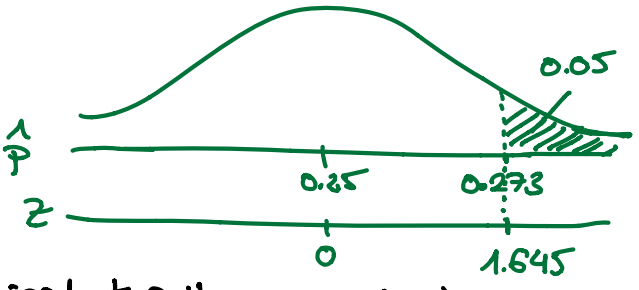
② Decide appropriate test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

③ Decide

SIGNIFICANCE LEVEL: α = = TYPE-I ERROR

one-tailed (right-tailed)
0.05



CRITICAL VALUE: $p = 0.273 \Rightarrow z = 1.645$

if our statistic is more extreme than this value ⇒ reject the H_0. = threshold value.

CRITICAL REGION: $= 0.05$

Make a decision

④ • Calculate test statistic from sample
• compare with sampling distribution
• Decide if you reject H_0

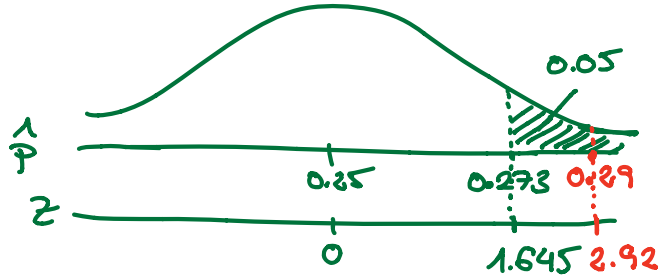
sample proportion $p = \frac{290}{1000} = 0.29$.

sampling distribution $p \sim N(p = 0.25, \sigma = \sqrt{\frac{0.25 \cdot 0.75}{1000}} = 0.01369)$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1) \Rightarrow z = \frac{0.29 - 0.25}{0.01369} =$$

$z = 2.92$

$2.92 > 1.645$
Reject H_0
Fail to reject H_1



ERRORS

• TYPE I ERROR: Reject H_0; H_0 true
α
• TYPE II ERROR: fail to reject H_0; H_0 false
β

⇒ ex: if pop. proportion $p = 0.27$.

$P(\beta) = P(\hat{p} < 0.273 \text{ if } p = 0.27) =$

$$P\left[z < \frac{0.273 - 0.27}{\sqrt{\frac{0.27 \cdot (0.73)}{1000}}}\right] = P(z < 0.214) = 0.5847$$

- SIGNIFICANCE TESTING = P-VALUE APPROACH.

ex: cancer patients : 90% die within 3 years.
new trt → reduced rate? $y = 128$ die " $\Rightarrow \hat{p} = \frac{128}{150} = 0.853$
$n = 150$ "
SIGNIFICANCE LEVEL : $\alpha = 0.05$

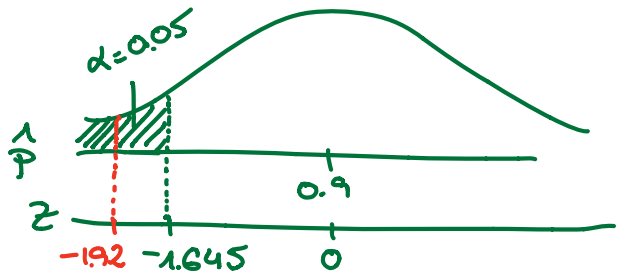
① STATE $\begin{cases} H_0 \\ H_1 \end{cases}$ ONE-TAILED TEST : $H_0: p = 0.90$ $H_1: p = < 0.90$ ONE TAIL!

② Decide appropriate test statistic

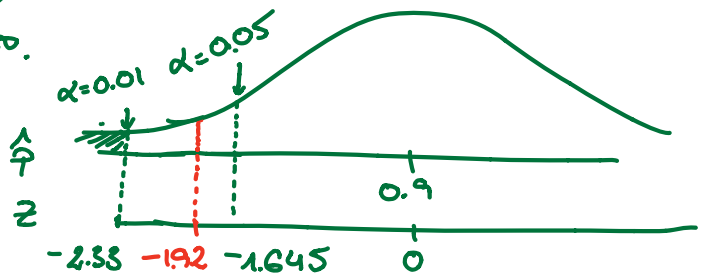
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.853 - 0.90}{\sqrt{\frac{0.90 \cdot 0.10}{150}}} = -1.92$$

③ Make a decision
- Calculate test statistic from sample
- compare with sampling distribution
- quantify how probable it is under H_0

$\alpha = 0.05$
$-1.92 < -1.645$
Reject the H_0.



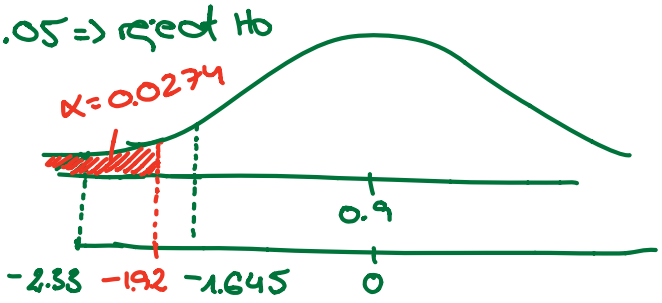
$\alpha = 0.01$
$-1.92 > -2.33$
Fail to Reject the H_0.



P-VALUE:
smallest significance level α that leads to us rejecting the H_0
or
the probability that we'd observe a more extreme statistic than we did if the H_0 were true

If p-value $< \alpha \Rightarrow$ Reject H_0
$P(\text{p-value} \leq \alpha \mid H_0) \leq \alpha \Rightarrow$ random vble

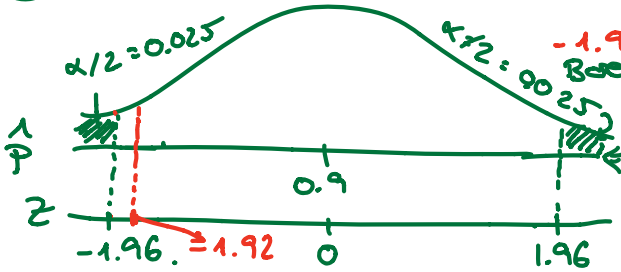
$0.0274 < 0.05 \Rightarrow$ reject H_0



p-value $\begin{cases} \text{The smallest } \alpha\text{-level that would lead to rejection of } H_0 \\ \alpha\text{-level associated with the test statistic } -1.92 \\ P(z < -1.92) = 0.0274. = \text{probability that we'd observe a } p = 0.853 \text{ if the true proportion were } 0.9 \ (= H_0 \text{ TRUE}) \end{cases}$

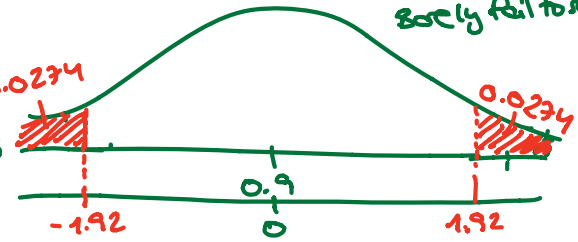
Ex: TWO-TAILED TEST : ① $H_0: p = 0.9$ ② $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.853 - 0.90}{\sqrt{\frac{0.90 \cdot 0.10}{150}}} = -1.92$
$H_1: p \neq 0.9$
SIGNIFICANCE LEVEL $\frac{\alpha}{2} = 0.025$

HYPOTHESIS TESTING :
③ CRITICAL VALUES : $-1.96, 1.96$ | Reject H_0 when : $|z| \geq 1.96 \rightarrow z > 1.96$ or $z < -1.96$
$-1.92 > -1.96$
Barely fail to reject H_0.



Insufficient evidence at $\alpha = 0.05$ level to conclude that the sample p differs significantly from $p = 0.9$

SIGNIFICANCE TESTING.
③ p-value $= P(|z| \geq 1.92) = P(z > 1.92 \text{ or } z < -1.92)$
$= 2 \times 0.0274 = 0.055 > 0.05$
Barely fail to reject H_0



TESTS ABOUT PROPORTIONS:

- ONE SAMPLE PROPORTION:

$Y \sim \text{binomial}(n, p)$ Y = nº of successes.

① $H_0: p = p_0$ $H_1: p \neq p_0$ => 2-SIDED TEST $\hat{p}$ is a probability of success = $\frac{Y}{n}$

$H_1: p > p_0$ or $H_1: p < p_0$ => 1-SIDED TEST.

② STATISTIC: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ when $n \to \infty$ due to the CLT (normal approximation of binomial)

A) Advantage: better approximation to α-level significance tests (p-value approach)

③ A) Reject H_0 if:

TWO-TAIL: $|z| = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/2}$. under H_0 $P(|Z| \geq z_{\alpha/2}) \approx \alpha$.

ONE-TAIL $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha$ or $z = \dfrac{\frac{Y}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_\alpha$

② B) ALTERNATIVE STATISTIC $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$

Advantage: the interpretation of the confidence interval is consistent with the hypothesis test decision. Also better estimate of the SD if the H_0 is clearly false.

③ B)

TWO-TAIL: $H_1: p \neq p_0$ = if p_0 is not in the $(1-\alpha)\cdot 100\%$ confidence interval

$z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \geq z_{\alpha/2}$ $\quad z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq -z_{\alpha/2}$

$\hat{p} - p_0 \geq z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\quad \hat{p} - p_0 \leq -z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$p_0 \geq \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\quad p_0 \leq \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Reject H_0 if:

ONE-TAIL: $H_1: p < p_0$ = if p_0 is not in the upper $(1-\alpha)\cdot 100\%$ confidence interval

$z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq -z_\alpha$

$\hat{p} - p_0 \geq -z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$p_0 \geq \hat{p} + z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \left[0, \hat{p} + z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

Ex: Manufacturer observes aprox $p = 0.06$ circuits fail.

New procedure → trials $Y_i \sim bin(200, p) \Rightarrow \hat{p} = \frac{Y}{200}$ is the new procedure better?

- C.I. approach: $\left[0, \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ → if it contains 0.06 → new procedure not necessarily better.
  if doesn't contain 0.06 → 95% confidence that the true p is less than 0.06. → new p. better.

- HYPOTHESIS approach

① $H_0: p = 0.06 = \hat{p}$ $H_1: \hat{p} < 0.06$ if $p = 0.06$, $Y = 12$ failures

② CRITICAL VALUE: new p. is better if $p = 0.035$, $Y = 7$ failures

Probability of $p = 0.06$ and observing 7 failures => TYPE I ERROR = α

Binomial => $\alpha = P(Y \leq 7; p = 0.06) = \sum_{y=0}^{7} \binom{200}{y} (0.06)^y (0.94)^{200-y} = 0.0829$

Poisson approximation => $\alpha \approx \sum_{y=0}^{7} \frac{12^y e^{-12}}{y!} = 0.090$ → low! because n small, p large

Prob. of an improvement to for ex. $p = 0.03$ and observing 7 failures => TYPE II ERROR = β

Binomial => $\beta = P(Y > 7, p = 0.03) = \sum_{y=8}^{200} \binom{200}{y} \cdot (0.03)^y \cdot (0.97)^{200-y}$

Poisson approx => $\beta \approx 1 - \sum_{y=0}^{7} \frac{6^y \cdot e^{-6}}{y!} = 1 - 0.744 = 0.256$ → high! ∴

$\frac{Y}{n} \sim N(\mu = p_0, \sigma^2 = \frac{p_0(1-p_0)}{n})$ when $n \to \infty$ under H_0 $Z \sim N(0,1)$

if $Z = \frac{\frac{Y}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha$ => $\frac{Y}{n}$ exceeds p_0 by z_α standard deviations of $\frac{Y}{n}$, we reject H_0.

The prob of $Z = \frac{\frac{Y}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha$ when H_0 is true $= \alpha$

significance level $= \alpha$

- TWO SAMPLE PROPORTIONS

① $H_0 : p_1 = p_2$ $\quad H_1 : p_1 - p_2 = 0$ or $p_1 \neq p_2$

$\hat{p} = \frac{y_1+y_2}{n_1+n_2}$: proportion of successes of the 2 samples combined.

② STATISTIC : $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

③

Reject H_0	2 TAIL : $H_1: p_1 \neq p_2$	ONE-TAIL $H_1: p_1 > p_2$	ONE-TAIL $H_1: p_1 < p_2$
	$\lvert z \rvert = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \geq z_{\alpha/2}$	$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > z_\alpha$	$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} < z_\alpha$

Proof : $\hat{p}_1 - \hat{p}_2 \sim N\left(\mu = p_1 - p_2, \sigma^2 = \left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)\right)$

if H_0 is true → $p_1 = p_2 = p \Rightarrow \sigma^2 = p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \overbrace{(p_1 - p_2)}^{0 \text{ bc } p_1 = p_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$$

we don't know the assumed common $p = p_1 = p_2 \Rightarrow$ pooled estimate $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$

Statistic : $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

②B) ALTERNATIVE STATISTIC : $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_1(1-\hat{p}_1)}{n_1} + \frac{p_2(1-\hat{p}_2)}{n_2}}}$

Advantage : the interpretation of the confidence interval is consistent with the hypothesis test decision

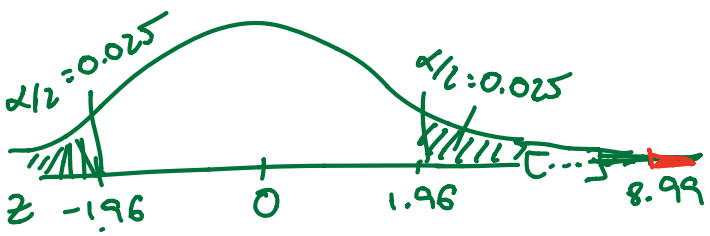
Ex: should cigarette tax be raised? y_i = "yes". Are the opinions between groups ≠?

	Non-smokers	Smokers
n	$n_1 = 605$	$n_2 = 195$
y	$y_1 = 351$	$y_2 = 41$
$\hat{p}$	$\hat{p}_1 = \frac{351}{605} = 0.58$	$\hat{p}_2 = \frac{41}{195} = 0.21$

$\hat{p} = \frac{41+351}{605+195} = 0.49$

① $H_0 = p_1 = p_2$
$H_1 = p_1 \neq p_2$

$z = \frac{(0.58 - 0.21) - 0}{\sqrt{0.49 \cdot (0.51)\left(\frac{1}{195}+\frac{1}{605}\right)}} = 8.99.$



$z > z_{\alpha/2} \Rightarrow$ Reject H_0.

$P = 2 \times P(z > 8.99) = 2(0.0000) = 0$

P-value < 0.0001

TESTS ABOUT MEANS

ASSUMPTION: sample pop { is $N \sim (\mu, \sigma)$; random sample }

- ONE SAMPLE MEAN

TWO-TAIL ONE-TAIL

① $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ or $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$

② STATISTIC

- KNOWN VARIANCES: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

- UNKNOWN VARIANCES: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$ if the data are normally distributed.

Ex: mean blood pressure in honolulu: $\mu = 120$ mmHg. $n = 100$. $\bar{x} = 130.1$. SD = 21.21 mmHg

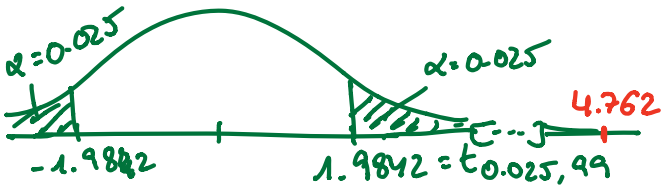
① $H_0: \mu = 120$ $H_1: \mu \neq 120$ ② $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{130.1 - 120}{21.2/\sqrt{100}} = 4.762$.

③ $\alpha = 0.05$

$t_{0.025,99} = 1.9842$

$t > t_{0.025,99}$

Reject H_0.



P-value $= 2 \times P(T_{99} > 4.762) < 2 \times P(T_{99} > 1.9842) = 2 \times 0.025 = 0.05$

Equivalent to C.I: $\bar{x} \pm t_{0.025,99}\left(\frac{S}{\sqrt{n}}\right) = 130.1 \pm 1.9842 \cdot \left(\frac{21.21}{\sqrt{100}}\right) = 130.1 \pm 4.21$

95% confident the honolulians blood pressure is between [125.89, 134.31]

- PAIRED T-TEST

① $H_0: \mu_x = \mu_y \Rightarrow \mu_x - \mu_y = \mu_D = 0$ $H_1: \mu_D \neq 0$

② STATISTIC: $\frac{\bar{d} - \mu_0}{S_D/\sqrt{n}} \sim t_{(n-1)} \Rightarrow \frac{\bar{d} - 0}{\frac{SD}{\sqrt{n}}}$

- TWO SAMPLE MEAN

ASSUMPTION: sample pop { are $N \sim (\mu, \sigma)$; random sample }

- UNKNOWN EQUAL VARIANCES:

① H_0: $\mu_x = \mu_y$ or $\mu_x - \mu_y = 0$ $\quad H_1$: $\mu_x - \mu_y \neq 0$, $\mu_x - \mu_y < 0$, $\mu_x - \mu_y > 0$

② STATISTIC: $T = \dfrac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(n+m-2)}$

POOLED 2-SAMPLE T-TEST

pooled sample variance.

$$s_p = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

PROOF:

$\bar{x} \sim N(\mu_x, \frac{\sigma^2}{n})$; $\bar{y} \sim N(\mu_y, \frac{\sigma^2}{m})$ $\Rightarrow$ $\bar{x} - \bar{y} \sim N(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ because x,y independent

$Z = \dfrac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0,1)$

$\dfrac{(n-1)s_x^2}{\sigma^2} \sim \chi^2_{n-1}$; $\dfrac{(m-1)s_y^2}{\sigma^2} \sim \chi^2_{m-1}$

bc x,y are normal $\Rightarrow$ $U = \dfrac{(n-1)s_x^2}{\sigma^2} + \dfrac{(m-1)s_y^2}{\sigma^2} \sim \chi^2_{n+m-2}$

$$T = \frac{Z}{\sqrt{\frac{U}{n+m-2}}} = \frac{\frac{1}{\sigma} \cdot \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\frac{1}{\sigma} \cdot \underbrace{\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{(n+m-2)}}}_{s_p}} = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

- UNKNOWN AND UNEQUAL VARIANCES:

① H_0: $\mu_x = \mu_y$ or $\mu_x - \mu_y = 0$ $\quad H_1$: $\mu_x - \mu_y \neq 0$, $\mu_x - \mu_y < 0$, $\mu_x - \mu_y > 0$

② Statistic: $T = \dfrac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \sim t_{(r)}$ $\longrightarrow$ $r = \dfrac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}$

WELCH'S T-TEST

use the integer [r]

same proof as ↓

WELCH'S t-interval: $\bar{x} - y \pm t_{\alpha/2, r} \cdot \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$

IF LARGE SAMPLE SIZE $Z = \dfrac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \sim N(0,1)$

Note: T and Z statistics do not do well if { . sample sizes, variances } severely different; . raw data is highly skewed

Then you need alternative methods.

Ex: Is the mean fastest speed driven by M ≠ than by Fem?

	n	mean	variance
MALE	$n = 34$	$\bar{x} = 105.5$	$s_x = 20.1$
FEMALE	$m = 29$	$\bar{y} = 90.9$	$s_y = 12.2$

Assume the 2 pops. are normally dist.

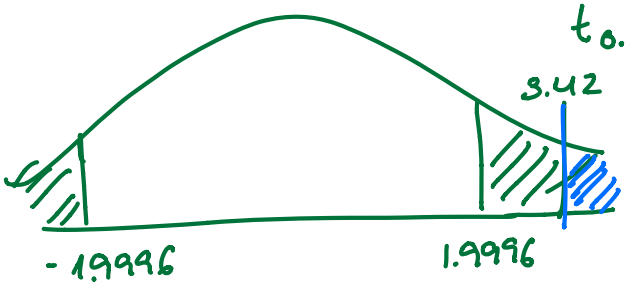
① $H_0: \mu_y = \mu_x$ $H_1: \mu_y \neq \mu_x$

EQUAL VARIANCES

pooled sample standard deviation

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} = \sqrt{\frac{33 \cdot (20.1)^2 + 28 \cdot (12.2)^2}{63-2}} = 16.9$$

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{(105.5 - 90.9) - 0}{16.9 \sqrt{\frac{1}{3} + \frac{1}{29}}} = 3.42.$$



$t_{0.025,61} = 1.9996$

$|t| > t_{0.025,61}$

Reject H_0.

P-value $= 2 \times P(T_{61} > 3.42) = 2(0.0006) = 0.0012$

NOT EQUAL VARIANCES.

$$r = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}} = \frac{\left(\frac{(20.1)^2}{34} + \frac{(12.2)^2}{29}\right)^2}{\frac{\left(\frac{(20.1)^2}{34}\right)^2}{33} + \frac{\left(\frac{(12.2)^2}{29}\right)^2}{28}}$$

$= 55.5$

$t = \ldots\ldots$

TEST ABOUT VARIANCES:

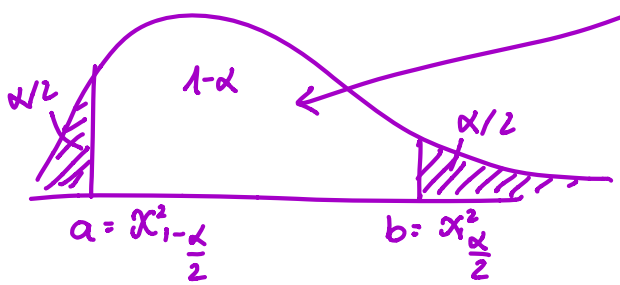
- ONE VARIANCE

① $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$ $\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$

② STATISTIC: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$

Proof

We know that if $X_i \sim N(\mu, \sigma^2)$ $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$



- 2 VARIANCES

① $H_0: \sigma_x^2 = \sigma_y^2 \simeq \frac{\sigma_x^2}{\sigma_y^2} = 1$ $\quad H_1: \sigma_x^2 \neq \sigma_y^2$ $\quad \sigma_x^2 > \sigma_y^2$ $\quad \sigma_x^2 < \sigma_y^2$

② STATISTIC : $F = \frac{S_x^2}{S_y^2}$

③ Reject H_0 : if F is too large $F > F_{\frac{\alpha}{2},(n-1,m-1)}$
if F is too small $F < F_{1-\frac{\alpha}{2},(n-1,m-1)} = \frac{1}{F_{\frac{\alpha}{2},(m-1,n-1)}}$

PROOF :

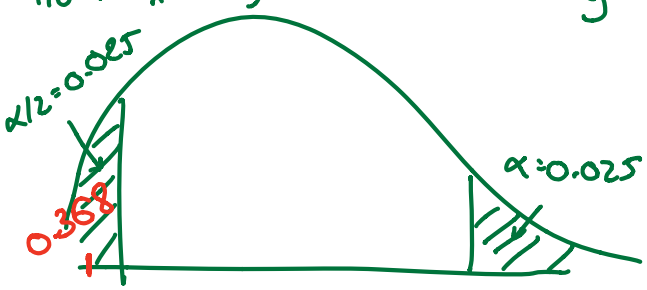
$X_i \sim N(\mu_x, \sigma_x^2) \Rightarrow \frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi^2_{n-1}$ $\quad Y_i \sim N(\mu_y, \sigma_y^2) \Rightarrow \frac{(m-1)S_y^2}{\sigma_y^2} \sim \chi^2_{m-1}$

ASSUMPTION : sample pop are $N \sim (\mu, \sigma)$, random sample

$$F = \frac{\frac{\frac{(m-1)S_y^2}{\sigma_y^2}}{m-1}}{\frac{\frac{(n-1)S_x^2}{\sigma_x^2}}{n-1}} = \frac{\sigma_x^2}{\sigma_y^2} \cdot \frac{S_y^2}{S_x^2} \sim F(m-1, n-1)$$

Ex: Is the variance of fastest speed driven by M ≠ than by Fem?

① $H_0: \sigma_x^2 = \sigma_y^2$ $\quad H_1: \sigma_x^2 \neq \sigma_y^2$ $\quad$ ② $F = \frac{(12.2)^2}{(20.1)^2} = 0.368.$



$F_{0.975(28,33)} = \frac{1}{F_{0.025(33,28)}} = \frac{1}{2.0880} = 0.478$ $\quad F_{0.025(28,33)} = 2.0441$

$F = 0.368 < 0.478$

Reject H_0 : two variances are not equal. So to test the difference between means, we should use Welch's t-test not the pooled 2-sample t-test.

- ONE or TWO -SIDED TEST ? Two-sided is more conservative : harder to reject H_0 hard to know directionality a priori
- Fail to reject H_0 vs accept H_0 : we don't know the H_0 is true, just that we don't have enough data to reject it. also acknowledges 0.05 random cut off.
- Don't forget assumptions ! they are actually implicitly included in the H_0.

CATEGORICAL . = NON-PARAMETRIC METHODS.

CHI-SQUARE .

- ONE SAMPLE :

CHI-SQUARE GOODNESS OF FIT STATISTIC : $Q_1 = \sum_{i=1}^{2} \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^{2} \frac{(\text{OBSERVED} - \text{EXPECTED})^2}{\text{EXPECTED}}$.

$$z^2 = \frac{(y_1 - np_1)^2}{np_1q_1} = \frac{(y_1 - np_1)^2}{np_1} + \frac{(y_2 - np_2)^2}{np_2}$$

PROOF : Ex : $y_1 + y_2 = n$; $p_1 + p_2 = 1$

y_1 = successes / females / ... $\quad y_1 \sim b(n, p_1) \rightarrow E(y_1) = np_1$, $\text{Var}(y_1) = np_1(1-p_1)$

y_2 = failures / males / .... $\quad y_2 = n - y_1 \sim b(n, p_2) = b(n, 1-p_1) \rightarrow E(y_2) = np_2 = n(1-p_1)$, $\text{Var}(y_2) = np_2(1-p_2) = n(1-p_1)(1-(1-p_1)) = np_1(1-p_1)$

when $n \to \infty$ $\begin{cases} np_1 \geq 5 \\ n(1-p_1) \geq 5 \end{cases} \xrightarrow{CLT} z = \frac{(y_1 - np_1)}{\sqrt{np_1(1-p_1)}} \sim N(0,1)$

$$z^2 = \frac{(y_1 - np_1)^2}{np_1(1-p_1)} = Q_1 \sim \chi^2_{(1)}$$

$$Q_1 = \frac{(y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(y_1 - np_1)^2}{np_1(1-p_1)} \cdot ((1-p_1) + p_1) = \frac{(y_1 - np_1)^2 \cdot (1-p_1)}{np_1(1-p_1)} + \frac{(y_1 - np_1)^2 p_1}{np_1(1-p_1)} =$$

$$\frac{(y_1 - np_1)^2}{np_1} + \frac{(y_1 - np_1)^2}{n(1-p_1)} = \frac{(y_1 - np_1)^2}{np_1} + \frac{(\overbrace{n - y_2}^{y_1 = n - y_2} - n(\overbrace{1 - p_2}^{p_1 = 1 - p_2}))^2}{np_2} = \frac{(y_1 - np_1)^2}{np_1} + \frac{(-(y_2 - np_2))^2}{np_2} =$$

$$= \frac{(y_1 - np_1)^2}{np_1} + \frac{(y_2 - np_2)^2}{np_2}$$

$\rightarrow Q_1 \sim \chi^2_{(1)}$: 1 df because if there is only 1 count y_1 $\rightarrow$ if you know y_1, you know $y_2 = n - y_1$

- we need CLT for proof so $\rightarrow$ $np_1 \geq 5$; $n(1-p_1) \geq 5$
- Q_1 is large if observed >> expected. Reject H_0 when Q_1 is large $\rightarrow$ depends on $\chi^2_{(1)}$

- For small samples ≤ 5
- very unequally distributed values among cells

} use FISCHER'S EXACT TEST

CHI-SQUARE is not adequate because it provides an approximation to the χ^2 that depends on a large sample size.

Ex: Rom state pop. 60% ♀, 40% ♂. sample $n=100$. $y_1 = ♀ = 53$ $y_2 = ♂ = 47$.
Is the sample random and representative of the population?

$H_0: p_F = 0.60$ $H_1: p_F \neq 0.60$. $\alpha = 0.05$. $\rightarrow \chi^2_{0.05,1} = 3.84$ Reject H_0 if $Q_1 \geq \chi^2_{0.05,1}$

$$z^2 = \frac{(y_1 - np_1)^2}{np_1q_1} = \frac{(y_1-np_1)^2}{np_1} + \frac{(y_2-np_2)^2}{np_2} = Q_1 = \frac{(53-60)^2}{60} + \frac{(47-40)^2}{40} = 2.04.$$

$Q_1 = 2.04 < \chi^2_{0.05,1} = 3.84 \Rightarrow$ fail to reject H_0.

RELATIONSHIP χ^2 and z: You can do a z test and square and it's the same!

$$z = \frac{0.53 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}} = -1.428 \Rightarrow z^2 = (-1.428) = 2.04. \quad |z| \geq 1.96 \Rightarrow Q_1 > (1.96)^2 > 3.84$$

$P = 2 \times P(z > 1.428) = 2(0.0766) = 0.1532 \Rightarrow P = P(\chi^2_{(1)} > 2.04) = 0.1532$

- K SAMPLES : K CATEGORIES : $Q_{K-1} = \sum_{i=1}^{K} \frac{(y_i - np_i)^2}{np_i} \sim \chi^2_{(K-1)}$

CATEGORIES	1	2	...	K-1	K
OBSERVED	y_1	y_2		y_{K-1}	$n - y_1 - y_2 - \dots - y_{K-1}$
EXPECTED	np_1	np_2		np_{K-1}	np_K

CONTINGENCY TABLES

i : n° of row categories $i = 1, \dots, h$
j : n° of col. categories $j = 1, \dots, K$

$\hat{p}_{ij}$: y_{ij}/n_i proportion of i^{th} sample in j^{th} category
$P(A_i \cap B_j)$ prob. that a randomly selected obs falls in $A_i \cap B_j$ cell

$\hat{p}_{\cdot j}$: $\sum_{i=1}^{h} p_{ij}$
$(y_{1j} + y_{2j})/n_1 + n_2$ proportion in j^{th} category
$p_{\cdot j} = P(B_j)$ prob that randomly selected obs falls in B_j column

$n_i = \sum_{j=1}^{K} p_{ij}$
$\sum_{j=1}^{K} y_{ij}$ n° in i^{th} sample
$p_{i\cdot} = P(A_i)$ prob that randomly selected obs falls in A_i row

	$j=1$	$j=2$	...	Total
$i=1$	y_{11} ($\hat{p}_{11}$)	y_{12} ($\hat{p}_{12}$)		$n_1 = \sum_{j=1}^{K} y_{1j} = p_{1\cdot}$
$i=2$	y_{21} ($\hat{p}_{21}$)	y_{22} ($\hat{p}_{22}$)		$n_2 = \sum_{j=1}^{K} y_{2j} = p_{2\cdot}$
⋮				⋮
Total	$y_{11}+y_{12}$ ($\hat{p}_1$)	$y_{12}+y_{22}$ ($\hat{p}_2$)	...	$n_1 + n_2 = n$
	($p_{\cdot 1}$)	($p_{\cdot 2}$)		

- TEST FOR HOMOGENEITY : are the proportions of B in different groups of A $=$ or ?
$\neq$

$H_0: p_{A_1 1} = p_{A_2 1}$; $p_{A_2 1} = p_{A_2 2}$ ...

CHI-SQUARE TEST STATISTIC: $Q = \sum_{i=1}^{h} \sum_{j=1}^{K} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}} \sim \chi^2_{(h-1)(K-1)}$

$h(K-1) - (K-1)$

Ex: Do Residents apply ↑ unnecessary blood transfusions ?

Residents $n = 71$; Attending physician $n = 49$. (%) % of row total p_{ij}/n_i

Physician	Frequent	Occasional	Rarely	Never	Total
Attending	2 (4.1%) $2/49 = 0.041$	3 (6.1%)	31 (63.3%)	13 (26.5%)	49
Resident	15 (21.1%)	28 (39.4%)	23 (32.4%)	5 (7.0%)	71
Total	17	31	54	18	120

$H_0: p_{RF} = p_{AF}$ AND $p_{RO} = p_{AO}$ .... $H_1: p_{RF} \neq p_{AF}$ OR $p_{RO} \neq p_{AO}$ ...

Physician	Frequent	Occasional	Rarely	Never	Total
Attending	$\frac{17}{120} \cdot 49 = 6.94$	$\frac{31}{120} \cdot 49 = 12.65$	$\frac{54}{120} \cdot 49 = 22.05$	7.35	49
Resident	10.058	18.342	31.95	10.65	71
Total	17	31	54	18	120

$Q = \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(2 - 6.94)^2}{6.94} + \ldots \frac{(5 - 10.65)^2}{10.65} = 31.88$

$\chi^2_{(4-1)(2-1)} = 3 : \chi^2_3 = 7.815$ $\quad Q > \chi^2_3 \Rightarrow 31.88 > 7.815 \rightarrow$ Reject H_0

Distribution of transfusions is different among physicians and residents

- TEST FOR INDEPENDENCE: are A and B independent from each other?

$H_0: P(A_i \cap B_j) = P(A_i) \times P(B_j)$

CHI-SQUARE TEST STATISTIC:

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{K} \frac{\left(y_{ij} - \frac{y_{i\cdot} y_{\cdot j}}{n}\right)^2}{\frac{y_{i\cdot} y_{\cdot j}}{n}} \sim \chi^2_{(h-1)(K-1)}$$

$h(K-1)-(K-1)$

Ex: Age and desire to ride bike independent?

Observed

	Age				
Wants to ride bike	18-24	25-34	35-49		
Yes	60	54			201
No	40	44			194
Total	100	98			395

Expected

	Age				
Wants to ride bike	18-24	25-34	35-49		
Yes	$\frac{100 \cdot 201}{395} = 50.88$	$\frac{98 \cdot 201}{395} = 49.86$			201
No					194
Total	100	98			395

$Q = \frac{(60-50.88)^2}{50.88} + \dots = 8.006 > \chi^2_3 = 7.815$ Reject H_0 → Riding bike depends on age!

MCNEMAR TEST: for 2×2 contingency table w/ dichotomous trait. or paired data.

	B^+	B^-	
A^+	a	b	a+b
A^-	c	d	c+d
	a+c	b+d	n

It tests consistency in responses across 2 vbles (NOT independence)

Do objects change from ⊕ to ⊖ randomly? we don't care about the subject who don't change (a, d)

$H_0: p_c = p_b$ $H_1: p_c \neq p_b$

TEST STATISTIC: $\chi^2 = \frac{(b-c)^2}{b+c}$

MANN-WHITNEY U TEST = WILCOXON RANK SUM TEST.

H_0: 2 populations are equal; H_1: 2 populations are not equal.

MANN-WHITNEY TEST STATISTIC: we always use the SMALLER of

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{n_2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{n_2} - R_2$$

R_1: sum of ranks in group 1

R_2: sum of ranks in group 2

Range of U: $0 - n_1 \times n_2$

complete separation → group 1 or 2 supports H_1

unlikely separation. supports H_0

Reject H_0 if $U \leq$ critical value

Ex:

		Total Sample (Ordered Smallest to Largest)		Ranks	
Placebo	New Drug	Placebo	New Drug	Placebo	New Drug
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	
SUM				37	18

$U_1 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 37 = 3$

$U_2 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 18 = 22$

for $\alpha = 0.05$ and $n_1 = 5, n_2 = 5$

$U = 2$ $U_1 > U \Rightarrow 3 > 2$

Fail to reject $H_0 \Rightarrow$ pop's are =

WILCOXON SIGNED RANK TEST

H_0: median difference = 0 H_1: median difference $\neq 0$ or > 0 or < 0

WILCOXON TEST STATISTIC: use the smallest

W_+ : Σ of $\oplus$ ranks

W_- : Σ of $\ominus$ ranks

If H_0 is true, we expect W_+ and W_- would be similar.

Reject H_0 if $W \leq$ critical value

Ex

Child	Before Treatment	After 1 Week of Treatment	Difference (Before-After)	Ordered Absolute Values of Difference Scores	Ranks	Signed Ranks
1	85	75	10	-5	1	-1
2	70	50	20	10	3	3
3	40	50	-10	-10	3	-3
4	65	40	25	10	3	3
5	80	20	60	15	5	5
6	75	65	10	20	6	6
7	55	40	15	25	7	7
8	20	25	-5	60	8	8

$W_+ = 34$

$W_- = 4$

Critical value of W for $n = 8$ $\alpha = 0.05$ = 6

$4 \leq 6 \Rightarrow$ reject $H_0 \Rightarrow$ There is a change

KRUSKAL-WALLIS TEST ≃ 1-WAY ANOVA w/ rank >2 groups

H_0: All the pop. medians are equal H_1: the pop medians ≠

TEST STATISTIC: $$H = \frac{12}{N(N+1)} \sum_{j=1}^{K} \frac{R_j^2}{n_j} - 3(N+1)$$

K: nº groups

N: total sample size

n_j: sample size jth group.

Reject H_0 if $H >$ critical value

Ex

			Total Sample (Ordered Smallest to Largest)			Ranks		
5% Protein	10% Protein	15% Protein	5% Protein	10% Protein	15% Protein	5% Protein	10% Protein	15% Protein
3.1	3.8	4.0	2.6			1		
2.6	4.1	5.5	2.9	2.9		2.5	2.5	
2.9	2.9	5.0	3.1			4		
	3.4	4.8		3.4			5	
	4.2			3.8			6	
					4.0			7
				4.1			8	
				4.2			9	
					4.8			10
					5.0			11
					5.5			12
SUM						7.5	30.5	40

$$H = \frac{12}{13 \cdot 13}\left[\frac{(7.5)^2}{3} + \frac{(30.5)^2}{5} + \frac{40^2}{4}\right] - 3 \cdot 13$$

$= 7.52$

H for $n_1 = 3, n_2 = 5, n_3 = 4$

$\alpha = 0.05$

is 5.656.

$7.52 > 5.656 \Rightarrow$ groups ≠

If there are 3 or more comparison groups and/or ≥5 obs. per group

$H \longrightarrow \chi^2_{K-1}$ so the critical value can be derived from the χ^2 table

FRIEDMAN TEST ≃ 1-WAY ANOVA w Rank and repeated measures

H_0:

TEST STATISTIC $$Q = \frac{12n}{K(K+1)} \sum_{j=1}^{K} \left(\bar{r}_{\cdot j} - \frac{K+1}{2}\right)^2$$

where $\bar{r}_{\cdot j} = \frac{\sum_{i=1}^{n} r_{ij}}{n}$

n: rows / blocks

K: columns / treatments

DERIVE α, β, POWER

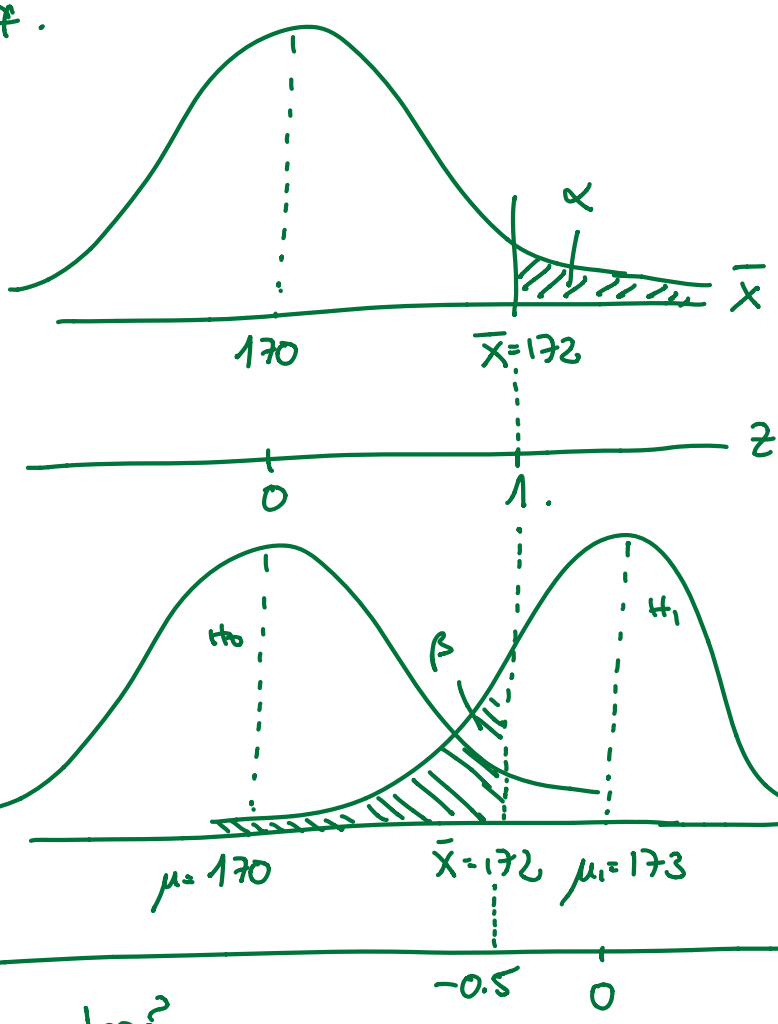
Ex: Random sample $X_i \sim N(\mu, \sigma = 10)$ $n = 25$. $H_0: \mu = 170$ $H_1: \mu > 170$.

If the agency decides $\bar{X} \geq 172$ cutoff.

what is the prob of Type I error?

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{172 - 170}{\frac{10}{\sqrt{25}}} = \frac{2}{2} = 1$$

$$P(\bar{X} > 172; \mu = 170) = P(Z \geq 1) = 0.1587$$



what if the true pop mean $\mu = 173$?

what is the prob of Type II error?

Fail to reject H_0 when H_1 is true

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{172 - 173}{\frac{10}{\sqrt{25}}} = -\frac{1}{2} = -0.5$$

$$P(\bar{X} < 172; \mu = 173) = P(Z < -0.5) = 0.3085$$

what is the prob the right decision is made (H_1 is accepted when it is true)?

$$P(\bar{X} \geq 172; \mu = 173) = P(Z \geq -0.5) = 0.6915 = 1 - 0.3085$$

	D^+ H_0	D^- H_1	
Fail to Reject H_0 T^+	a $D^+ \cap T^+$	b $D^- \cap T^+$	$T^+ = a+b$
reject H_0 T^-	c $D^+ \cap T^-$	d $D^- \cap T^-$	$T^- = c+d$
	D^+ $a+c$	D^- $b+d$	

$$se: \frac{D^+ \cap T^+}{D^+} = \frac{a}{a+c}$$

$$sp: \frac{D^- \cap T^-}{D^-} = \frac{d}{b+d}$$

$$1 - se: \alpha = P(H_1 | H_0) = \frac{c}{a+c} = \text{False} \ominus$$

$$1 - sp: \beta = P(H_0 | H_1) = \frac{b}{b+d} = \text{False} \oplus$$

POWER FUNCTIONS

Power: the probability of making the correct decision if the H_1 is true → reject the H_0.

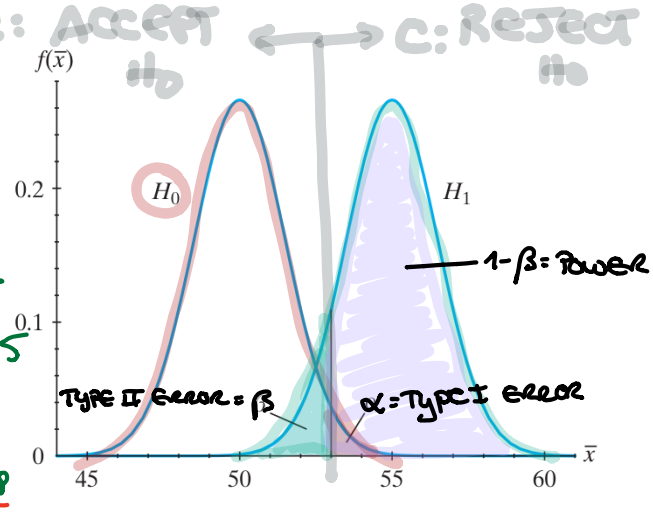


Figure 8.1-1 pdf of $\overline{X}$ under H_0 and H_1

Ex: $X_i \sim N(\mu, \sigma^2 = 16)$ $n = 16$ students. $\alpha = 0.05$.

① $H_0: \mu = 100$; $H_1: \mu > 100$ ② z statistic = 1.645

③ Reject H_0 if $z \geq 1.645$ or $\mu \geq 106.58$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow x = \mu + z \cdot \frac{\sigma}{\sqrt{n}} = 100 + 1.645\left(\frac{16}{\sqrt{16}}\right) = 106.58$$

H_0 Distribution: centered around 0.

H_1 Distribution: not "

↳ NON CENTRALITY DISTRIBUTION.

- True pop. $\mu = 108$

Power = $P(\bar{x} \geq 106.58 \mid \mu = 108) =$

$$P\left(z \geq \frac{106.58 - 108}{16/\sqrt{16}}\right) = P(z \geq -0.36) =$$

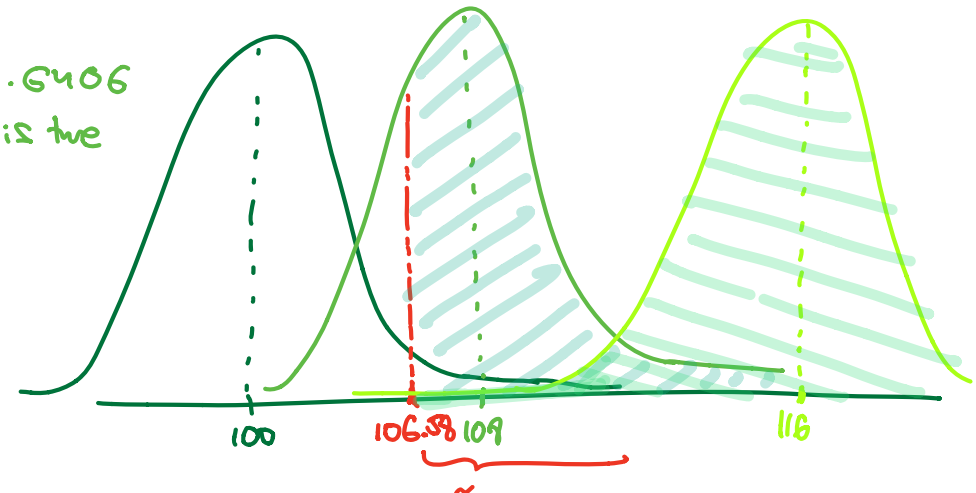
$= 1 - P(z < -0.36) = 1 - \Phi(-0.36) = 0.6406$

64% chance of rejecting H_1 when H_1 is true

- True pop $\mu = 116$

Power = $P(\bar{x} \geq 106.58 \mid \mu = 116) =$

$P(z \geq -2.36) = 0.9909$



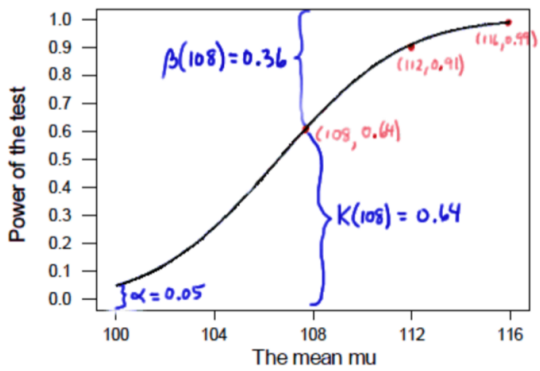
Power ↑ the further the actual μ moves away from the H_0.

Power = $1 - \Phi(z)$ where $z = \frac{106.58 - \mu}{16/\sqrt{16}}$

POWER FUNCTION:

$$K(\mu) = 1 - \Phi\left(\frac{\text{CRITICAL VALUE} - \mu}{s/\sqrt{n}}\right)$$

The power will depend on the the value of μ



Add sample size calculator ??

BEST CRITICAL REGIONS

Recall:

LIKELIHOOD FUNCTION : $L(\theta)$ (for 1 parameter)

Random sample : $X_1, X_2 \dots X_n$. X_i is independent bc it's random sample

the pdf for each X_i is $f(x_i, \theta)$

Joint probability mass function of X_i independent vbles

$$L(\theta) = P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \dots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

COMPOSITE HYPOTHESIS : A hypothesis that does not uniquely specify the distribution of the pop. from which the sample is taken.

Ex : $X_i \sim \exp(\theta)$. $H : \theta > 2$ $\quad f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \Rightarrow \frac{1}{\theta} e^{-\frac{x}{3}}$ or $\frac{1}{\theta} e^{-\frac{x}{22}}$ ...etc

$X_i \sim N(\mu, \sigma^2)$. $H : \mu = 12$ $\quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[\frac{-(x-12)^2}{2\sigma^2}\right]$ → μ defined, σ^2 not defined.

BEST CRITICAL REGION :

$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_1$

C and D are critical regions of size α → $\alpha = P(C; \theta_0)$, $\alpha = P(D; \theta_0)$

C is the best critical region of size α if $P(C; \theta_0) \geq P(D; \theta_0)$

that is if the power of C is at least as great as the power of every other critical region D of size α. = the power of the test at $\theta = \theta_0$ is largest among all possible hypothesis tests.

- NEYMAN-PEARSON LEMMA H_0: simple H ; H_1: simple H

$X_i \sim$ prob dist (θ) . X_i: random sample -

If C is a critical region of size α and K is a constant such that:

$\frac{L(\theta_0)}{L(\theta_\alpha)} \leq K$ inside critical region C

$\frac{L(\theta_0)}{L(\theta_\alpha)} \geq K$ outside critical region C

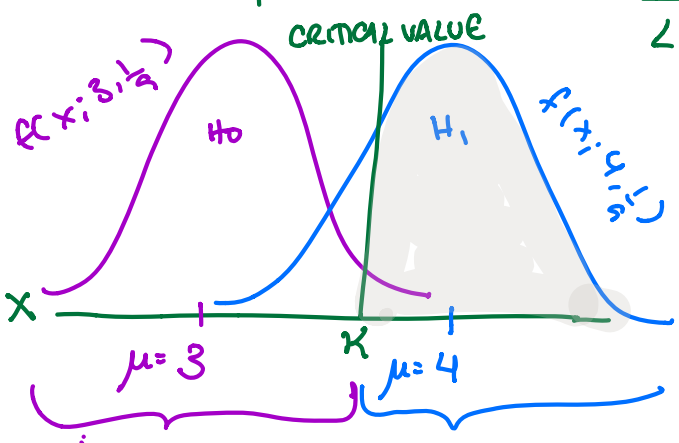
} then C is the best critical region for testing the simple $H_0: \theta = \theta_0$ against the simple $H_1: \theta \neq \theta$

Ex: X is one data point from $X_i \sim N(\mu, \sigma = \frac{1}{3})$ $H_0: \mu = 3$ vs $H_1: \mu = 4$

The most powerful test $\Rightarrow \frac{L(\mu_0)}{L(\mu_\alpha)} = \frac{L(3)}{L(4)} = \frac{f(x; 3, 1/9)}{f(x; 4, 1/9)}$

→ small for some points x inside critical region $\leq K$

↓ Large for some points x outside critical region $\geq K$



we reject $H_0: \mu = 3$ if our observed x is large in favor of $H_1: \mu = 4$, ≃ if x falls in critical region C

$\frac{f(x; 3, 1/9)}{f(x; 4, 1/9)}$ Large $\frac{f(x; 3, 1/9)}{f(x; 4, 1/9)}$ small

Ex: X is a single observation from $X_i \Rightarrow pdf = f(x) = \theta x^{\theta-1}$ $0 < x < 1$

Best critical region with $\alpha = 0.05$ to test $H_0: \theta = 3$ $H_1: \theta = 2$ → simple hypotheses

$\frac{L(\theta_0)}{L(\theta_\alpha)} = \frac{3x^{3-1}}{2x^{2-1}} = \frac{3x^2}{2x} = \frac{3}{2}x \leq K \Rightarrow x < \frac{2}{3}K = K^* \Rightarrow$ rejection region for the most powerful test

Find K^*! we want $\alpha = P(\text{Type I Error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = 0.05$

$\alpha = P(X < K^* \text{ when } \theta = 3) = \int_0^{K^*} 3x^2 dx = 0.05 \Rightarrow 3 \cdot \frac{1}{3} x^3 \Big|_0^{K^*} = (K^*)^3 = 0.05$

$K^* = (0.05)^{1/3} = 0.368$

Neyman-Pearson lemma says the rejection region for the most powerful test to test $H_0: \theta = 3$ $H_1: \theta = 2$ is $x < 0.368$.

Ex: X_i random sample $\sim N(\mu, \sigma^2 = 16)$ Best critical region for $\begin{cases} n = 16 \quad \alpha = 0.05 \\ \text{simple: } H_0: \mu = 10; \\ \text{simple: } H_1: \mu = 15 \end{cases}$

$L(\theta)$ for $\sim N(\mu, \sigma^2)$: $(2\pi\sigma_0^2)^{-\frac{n}{2}} \cdot \exp\left[\frac{-1}{2\sigma_0^2}\sum(x_i-\mu_0)^2\right]$

$$\frac{L(10)}{L(15)} = \frac{(32\pi)^{-16/2} \cdot \exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i-10)^2\right]}{(32\pi)^{-16/2} \cdot \exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i-15)^2\right]} = e^{\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i-10)^2\right] - \left[-\frac{1}{32}\sum_{i=1}^{16}(x_i-15)^2\right]}$$

$$= e^{-\frac{1}{32}\left[\sum_{i=1}^{16}(x_i-10)^2 - \sum_{i=1}^{16}(x_i-15)^2\right]} = e^{-\frac{1}{32}\left[\sum x_i^2 - 2\cdot 10\sum x_i + \sum_{i=1}^{16}100 - \left(\sum x_i^2 - 2\cdot 15\cdot\sum x_i + 225\right)\right]}$$

($20\sum x_i$, 160, $30\sum x_i$)

take ln on both sides

$$= e^{-\frac{1}{32}\left[10\sum x_i + 1600 - 3600\right]} = e^{-\frac{1}{32}(10\sum x_i - 2000)} \leq K \Rightarrow -10\sum x_i + 2000 \leq \ln K \cdot 32$$

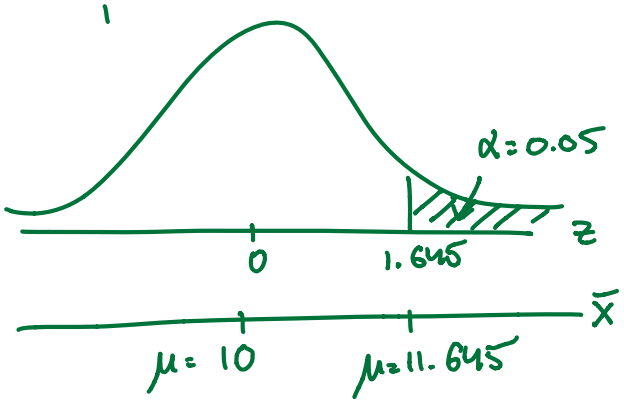
$$\Rightarrow \frac{-10}{-n\cdot 10}\sum x_i \leq \frac{(32\cdot\ln K - 2000)}{-n\cdot 10} \Rightarrow \frac{\sum x_i}{n} \geq \frac{(32\cdot\ln K - 2000)}{160} \Rightarrow \bar{x} \geq K^*$$

Neyman Pearson: Critical region $C = [(x_1, x_2 \dots x_n); \bar{x} \geq K^*]$ where K^* is selected so that the size of the critical region $C = \alpha = 0.05$?

under the H_0 the sample mean has distribution $\bar{x} = 10$, $SD = \frac{\sigma}{\sqrt{n}} = \frac{4}{4} = 1 \Rightarrow \bar{X} \sim N(10, 1)$

$$P[\bar{X} \geq K^*; \mu = 10] = P\left[\frac{\bar{X}-10}{1} \geq \frac{K^*-10}{1}; \mu = 10\right] = 1 - P\left(z < \frac{K^*-10}{1}\right) = 0.05$$

$\frac{K^*-10}{1} = 1.645 \Rightarrow K^* = 11.645$



The rejection region for the most powerful test to test $H_0: \mu = 10$; $H_1: \mu = 15$ is

$\bar{x} \geq 11.645$

The power of this test when $\mu = 15$

$$P(\bar{X} > 11.645; \mu = 15) = P\left(z > \frac{11.645 - 15}{\frac{\sqrt{16}}{\sqrt{16}}}\right) =$$

$P(z > -3.36)$

↓

this is the max power we'd set at of any test

- UNIFORMLY MOST POWERFUL TESTS = UMP H_0: simple H. H_1: composite H?

A test defined by the critical region C for a size α is the UMP test if it is the most powerful test against each simple alternative in the H_1.
The critical region C = UMP critical region for size α.

Ex: X: random sample $\sim N(\mu, \sigma^2=16)$ best critical region for $\begin{cases} n=16 \quad \alpha=0.05 \\ \text{simple: } H_0: \mu=10, \\ \text{composite: } H_1: \mu>10 \end{cases}$

$L(\theta)$ for a $N(\mu,\sigma^2)$: $(2\pi\sigma_0^2)^{-\frac{n}{2}} \cdot \exp\left[\frac{-1}{2\sigma_0^2}\sum(x_i-\mu_0)^2\right]$

$$\frac{L(10)}{L(\mu_\alpha)} = \frac{(32\pi)^{-16/2}\cdot \exp\left[-\frac{1}{32}\cdot\sum_{i=1}^{16}(x_i-10)^2\right]}{(32\pi)^{-16/2}\cdot \exp\left[-\frac{1}{32}\cdot\sum_{i=1}^{16}(x_i-\mu_\alpha)^2\right]} = e^{-\frac{1}{32}\left[\sum_{i=1}^{16}(x_i-10)^2-\sum_{i=1}^{16}(x_i-\mu_\alpha)^2\right]} \leq K$$

take ln on both sides

$$e^{-\frac{1}{32}\left[\sum x_i^2 - 2\cdot 10\sum x_i + 16(10)^2 - \left(\sum x_i^2 - 2\cdot\mu_\alpha\sum x_i - 16\mu_\alpha^2\right)\right]} = e^{-\frac{1}{32}\left(-2\cdot\sum x_i(10-\mu_\alpha)+16(\mu_\alpha^2-10^2)\right)} \leq K$$

(160 = $20\sum x_i$; $30\sum x_i$)

more constants to right

$$-2\cdot\sum x_i(10-\mu_\alpha)+16(\mu_\alpha^2-10^2) \leq 32\ln(K) \Rightarrow \frac{\sum x_i}{-n} \geq \frac{32\cdot\ln(K)-16(\mu_\alpha^2-10^2)}{2\cdot(10-\mu_\alpha)\cdot -n}$$

$$\bar{x} \geq -\frac{32\cdot\ln(K)-16(\mu_\alpha^2-10^2)}{16\cdot 2\cdot(10-\mu_\alpha)} = K^* \Rightarrow \frac{L(10)}{L(\mu_\alpha)} \leq K \Rightarrow \bar{x} \geq K^*$$

Neyman Pearson: The best critical region to test $H_0: \mu=10$ against each $H_1: \mu=\mu_1$

where $\mu_1 > 10$ is $C = [(x_1, x_2 \dots x_n);\ \bar{x} \geq K^*]$

where K^* is selected so that the size of the critical region $C = \alpha = P(\text{Type I Error})$

$\alpha = P(\bar{X} \geq K^*)$ when $\mu = 10$

LIKELIHOOD RATIOS

H_0: simple / composite H H_1: simple / composite H

Ω : total possible parameter space of Θ in both the H_0 and H_1

$H_0: \Theta \in \omega$ H_0 where ω is a subset of Ω

$H_1: \Theta \in \omega'$ H_1 where ω' is the complement of ω in space Ω.

$L(\hat{\Omega})$: maximum likelihood function with respect to Θ when Θ is in entire space Ω

$L(\hat{\omega})$: maximum likelihood function with respect to Θ when Θ is in null space ω H_0.

LIKELIHOOD RATIO : $\lambda = \dfrac{L(\hat{\omega})}{L(\hat{\Omega})}$

In general, a LR is simply the likelihood of one hypothesis relative to that of another, for an observed set of data. This is, calculated as the probability of the data given one hypothesis, often defined by a particular value of a parameter, divided by the probability of the same data given the second, competing hypothesis.

CRITICAL REGION FOR THE LIKELIHOOD RATIO TEST to test $H_0: \Theta \in \omega$ against $H_1: \Theta \in \omega'$ is the set of sample points for which $\lambda = \dfrac{L(\hat{\omega})}{L(\hat{\Omega})} \leq K$ where $0 < K < 1$ and K is selected so that the test has a desired significance α

- H_0: simple H H_1: composite H

Ex: Volume of honey in a jar 10 oz. $x_i \sim N(\mu = 10, \sigma^2 = 2)$ Likelihood ratio test for testing $H_0: \mu = 10$ vs. $H_1: \mu \neq 10$ at $\alpha = 0.05$?

$\Omega = \{\mu : -\infty < \mu < \infty\}$; H_0: $\omega = \{\mu = 10\}$

$L(\hat{\omega}) = L(10)$ when H_0 is true.

$L(\hat{\Omega})$ => what value of μ maximizes $L(\mu)$ => MLE of μ => MLE of $\hat{\mu} = \bar{x}$ =>

$L(\hat{\Omega}) = L(\bar{x})$

$L(\theta)$ for a $N(\mu, \sigma^2)$: $(2\pi\sigma_0^2)^{-\frac{n}{2}} \cdot e^{\left[-\frac{1}{2\sigma_0^2}\sum (x_i-\mu_0)^2\right]}$

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{L(10)}{L(\bar{x})} = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{2}} \cdot e^{\left[-\frac{1}{2\cdot 2}\sum (x_i-10)^2\right]}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{2}} \cdot e^{\left[-\frac{1}{2\cdot 2}\sum (x_i-\bar{x})^2\right]}}$$

$$\left[-\frac{1}{4}\cdot \sum (x_i - \bar{x} + \bar{x} - 10)^2\right] = -\frac{1}{4}\left[\sum (x_i-\bar{x})^2 + 2(\bar{x}-10)\underbrace{\sum (x_i-\bar{x})}_{0} + \underbrace{\sum (\bar{x}-10)^2}_{n(\bar{x}-10)^2}\right] = -\frac{1}{4}\left[\sum (x_i-\bar{x})^2 + n(\bar{x}-10)^2\right]$$

$$= \frac{\exp\left[-\frac{1}{4}\sum (x_i-\bar{x})^2\right] \exp\left[-\frac{n}{4}(\bar{x}-10)^2\right]}{\exp\left[-\frac{1}{4}\sum (x_i-\bar{x})^2\right]} \Rightarrow \lambda = e^{\left[-\frac{n}{4}(\bar{x}-10)^2\right]} \leq K$$

repeat the H_0 when λ is small (less than some K)

$$-\frac{n}{4}(\bar{x}-10)^2 \le \ln K \Rightarrow (\bar{x}-10)^2 \ge -\ln K \cdot \frac{4}{n} \Rightarrow \frac{|\bar{x}-10|}{\sigma/\sqrt{n}} \ge \frac{\sqrt{-\frac{4}{n}\cdot \ln K}}{\sigma/\sqrt{n}} = K^*$$

$\downarrow$

$Z \sim N(0,1)$ for $H_0: \mu = 10$

λ ratio tells us to reject $H_0: \mu = 10$ in favor of $H_1: \mu \neq 10$ for all sample means for which $\frac{|\bar{x}-10|}{\sigma/\sqrt{n}} \ge z_{0.025} = 1.96$ is true.

↳ because we set α at 0.05

- H_0: composite H H_1: composite H

Ex: $X \sim N(\mu, \sigma^2)$

Likelihood ratio test for testing $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ at $\alpha = 0.05$?

$\Omega = \{(\mu, \sigma^2): -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$; $\omega = \{(\mu, \sigma^2): \mu = \mu_0, 0 < \sigma^2 < \infty\}$.

MLE for $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$L(\hat{\Omega}) = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \cdot \exp\left[\frac{-1}{2\hat{\sigma}^2}\sum(x_i-\hat{\mu})^2\right] = \left[\frac{1}{2\pi \underbrace{\frac{1}{n}\sum(x_i-\bar{x})^2}_{\hat{\sigma}^2}}\right]^{\frac{n}{2}} \cdot \exp\left[-\frac{\sum(x_i-\bar{x})^2}{2\frac{1}{n}\sum(x_i-\bar{x})^2}\right]$$

$$= \left[\frac{n}{2\pi\sum(x_i-\bar{x})^2}\right]^{\frac{n}{2}} \cdot e^{-\frac{n}{2}} = \left[\frac{n\cdot e^{-1}}{2\pi\sum(x_i-\bar{x})^2}\right]^{\frac{n}{2}}$$

under the H_0 parameter space ω MLE for $\hat{\mu} = \mu_0$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$

$$L(\hat{\omega}) = \left[\frac{n\cdot e^{-1}}{2\pi\sum(x_i-\mu_0)^2}\right]^{\frac{n}{2}}$$

$$\lambda = \frac{L(\hat{\Omega})}{L(\hat{\omega})} = \frac{\left[\frac{n e^{-1}}{2\pi\sum(x_i-\bar{x})^2}\right]^{\frac{n}{2}}}{\left[\frac{n e^{-1}}{2\pi\sum(x_i-\mu_0)^2}\right]^{\frac{n}{2}}} = \left[\frac{2\pi\sum(x_i-\bar{x})^2}{2\pi\sum(x_i-\mu_0)^2}\right]^{\frac{n}{2}} = \left[\frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2}\right]^{\frac{n}{2}} = \left[\frac{1}{1+\frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2}}\right]^{\frac{n}{2}}$$

$$\frac{1}{1+\frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2}} \le K^{\frac{2}{n}} \Rightarrow \frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2} \cdot \frac{1}{\frac{1}{n-1}} \ge (n-1)(K^{-2/n}-1)$$

when H_0 is true : $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$ $\quad \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$

$$H_0: T = \frac{\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}}{\sqrt{\frac{\sum_{i=1}^{n}(\bar{X}_i-\bar{X})^2}{\sigma^2(n-1)}}} = \frac{n(\bar{x}-\mu_0)^2}{\frac{1}{n-1}\sum(x_i-\bar{x})^2} = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \sim t_{\frac{\alpha}{2},n-1} = K^*$$

$t_{\alpha/2,n-1}$ is the size α t-test that leaves the prob of committing type I error $= \alpha$

(1) STATISTIC for the LIKELIHOOD RATIO : $-2\log LR$

under regularity conditions on the model $f(x|\theta)$:
- $\theta \in$ in H_0 parameter space
- $n \to \infty$

$LR: -2\log R \longrightarrow \chi^2_{df}$ → Difference in the number of parameters
Number of free parameters specified by θ in the H_0 space
− Number of parameters in the entire parameter space.

(2) Reject H_0 if $-2\log R > \chi^2_{df,\alpha}$

ANOVA. ANALYSIS OF VARIANCE

- ONE-WAY ANOVA : categorical variables

m independent random variables

m normal distributions $\begin{cases} \text{means}: \mu_1, \mu_2 \dots \mu_m \\ \text{variance}: \sigma^2 \end{cases}$

$X_{ini} \sim N(\mu_i, \sigma^2)$ $i = 1 \dots m$

① $H_0: \mu_1 = \mu_2 \dots = \mu_m = \mu$

H_1: NOT H_0 At least 1 mean differ

Table 9.3-1 One-factor random samples

Group	Data : j observations				Means
X_1:	X_{11}	X_{12}	$\cdots$	X_{1n_1}	$\overline{X}_{1\cdot}$
X_2:	X_{21}	X_{22}	$\cdots$	X_{2n_2}	$\overline{X}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_m:	X_{m1}	X_{m2}	$\cdots$	X_{mn_m}	$\overline{X}_{m\cdot}$
Grand Mean:					$\overline{X}_{\cdot\cdot}$

Table 9.3-2 Analysis-of-variance table

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F Ratio
Treatment	SS(T) BETWEEN	$m-1$	$\text{MS(T)} = \frac{\text{SS(T)}}{m-1}$	$\frac{\text{MS(T)}}{\text{MS(E)}}$
Error	SS(E) WITHIN	$n-m$	$\text{MS(E)} = \frac{\text{SS(E)}}{n-m}$	
Total	SS(TO)	$n-1$		

P-value $= P\left(F_{(m-1,\,n-m)} \geq F = \frac{\text{MS(T)}}{\text{MS(E)}}\right)$

↓ unbiased estimator of σ^2

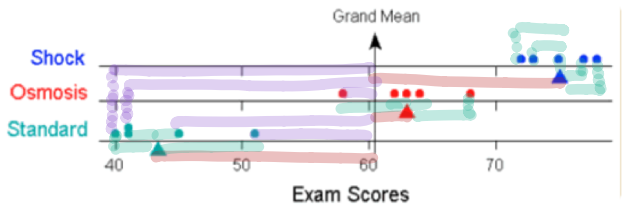
m : nº groups compared

x_{ij}: jth obs of ith group . $j = 1 \dots n_i$

the nº of obs may vary per group

$\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{i=1}^{n_i} x_{ij}$ mean of group i

$\bar{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{m} \sum_{i=1}^{n_i} x_{ij}$ grand mean of all data points



ASSUMPTIONS

$SS(T) = \sum_{i=1}^{m} \sum_{i=1}^{n_i} (\bar{x}_i - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n_i \bar{X}_i^2 - n\bar{X}_{\cdot\cdot}^2 \Rightarrow \frac{SST}{\sigma^2} \sim \chi^2_{(m-1)}$ ← H_0 is true

Difference group mean vs grand mean

$SS(E) = \sum_{i=1}^{m} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2 = SS(TO) - SS(T) \Rightarrow \frac{SSE}{\sigma^2} \sim \chi^2_{(n-m)}$

- $x_{ij} \sim N(\mu_i, \sigma^2)$
- $S_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}$ sample variance x_{ij}

Difference observations vs group mean

Random Error Treatment

$SS(TO) = \sum_{i=1}^{m} \sum_{i=1}^{n_i} (x_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{m} \sum_{i=1}^{n_i} (x_{ij}^2 - n\bar{X}_{\cdot\cdot}^2) = \overbrace{SS(E)} + \overbrace{SS(T)} \Rightarrow \frac{SS(TO)}{\sigma^2} \sim \chi^2_{(n-1)}$

Difference observations vs. grand mean.

If $X_{ij} \sim N(\mu_i, \sigma^2) \Rightarrow F = \dfrac{\frac{SST}{\sigma^2}/(m-1)}{\frac{SSE}{\sigma^2}/(n-m)} = \frac{MST}{MSE} \sim F_{(m-1,\,n-m)}$

compares the "average" variability between groups to "average" variability within groups

F is a reasonable statistic because

$E(MSE) = \sigma^2$

under H_0 ($\mu_1 = \mu_2 = \dots = \mu_m$) $E(MST) = \sigma^2 \Rightarrow \frac{MST}{MSE} \approx \frac{\sigma^2}{\sigma^2} \sim 1$

under H_1 $E(MST) > \sigma^2 \Rightarrow \frac{MST}{MSE} > 1$

(2) Reject the H_0 if $F \geq F_\alpha(m-1, n-m)$ or $P = P(F(m-1, n-m) \geq F) \leq \alpha$

ASSUMPTIONS of F-test: { independence, normality, equality of means

Ex.:

Table 9.3-3 Illustrative data

	Observations			$\bar{X}_{i\cdot}$
x_1:	13	8	9	10
x_2:	15	11	13	13
x_3:	8	12	7	9
x_4:	11	15	10	12
$\bar{x}_{\cdot\cdot}$				11

Let X_1, X_2, X_3, X_4 be independent random variables that have normal distributions $N(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4$. We shall test

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

against all alternatives on the basis of a random sample of size $n_i = 3$ from each of the four distributions. A critical region of size $\alpha = 0.05$ is given by

$$F = \frac{\text{SS(T)}/(4-1)}{\text{SS(E)}/(12-4)} \geq 4.07 = F_{0.05}(3, 8).$$

Table 9.3-4 ANOVA table for illustrative data

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F Ratio	p-value
Treatment	30	3	30/3	1.6	0.264
Error	50	8	50/8		
Total	80	11			

For the given data, the calculated SS(TO), SS(E), and SS(T) are

$$\text{SS(TO)} = (13-11)^2 + (8-11)^2 + \cdots + (15-11)^2 + (10-11)^2 = 80,$$

$$\text{SS(E)} = (13-10)^2 + (8-10)^2 + \cdots + (15-12)^2 + (10-12)^2 = 50,$$

$$\text{SS(T)} = 3[(10-11)^2 + (13-11)^2 + (9-11)^2 + (12-11)^2] = 30.$$

Note that since SS(TO) = SS(E) + SS(T), only two of the three values need to be calculated directly from the data. Here the computed value of F is

$$\frac{30/3}{50/8} = 1.6 < 4.07,$$

and H_0 is not rejected. The p-value is the probability, under H_0, of obtaining an F that is at least as large as this computed value of F. It is often given by computer programs.

MULTIPLE COMPARISONS: once you have rejected the H_0, how do you know which means differ from each other?

- Exploratory analysis: sort means by magnitude, box-whisker plot.
- Testing: Tuckey's honest significant difference test, etc --

- TWO-WAY ANOVA

- ONE VALUE PER CELL, NO REPLICATES $\Rightarrow$ $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ independent rvs. $i = 1, 2, \dots$ a → row; $j = 1, 2, \dots$ b → col. a×b = n comb.

Table 9.4-1 Two-way ANOVA table, one observation per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \frac{\text{SS(A)}}{a-1}$	$\frac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \frac{\text{SS(B)}}{b-1}$	$\frac{\text{MS(B)}}{\text{MS(E)}}$
Error	SS(E)	$(a-1)(b-1)$	$\text{MS(E)} = \frac{\text{SS(E)}}{(a-1)(b-1)}$	
Total	SS(TO)	$ab-1$		

$\mu_{ij} = \mu + \alpha_i + \beta_j$ (↓ ith row effect, ↓ jth col effect, + overall effect.) $\quad \sum_{i=1}^{a} \alpha_i = 0 = \sum_{j=1}^{b} \beta_j = 0$

H_0: $H_A: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$; $H_B: \beta_1 = \beta_2 = \dots = \beta_b = 0$ $\Rightarrow$ if true: $\frac{SS(A)}{\sigma^2} \sim \chi^2_{(a-1)}$; $\frac{SS(B)}{\sigma^2} \sim \chi^2_{(b-1)}$; $\frac{SS(E)}{\sigma^2} \sim \chi^2_{(a-1)(b-1)}$

$$SS(TO) = \sum_{i=1}^{a}\sum_{j=1}^{b} (x_{ij} - \bar{x}_{..})^2 = SS(A) + SS(B) + SS(E) \sim \chi^2_{(ab-1)}$$

Test $H_A \Rightarrow F = \frac{MS(A)}{MS(E)}$ Reject H_A when F is too large

Test $H_B \Rightarrow F = \frac{MS(B)}{MS(E)}$ Reject H_B when F is too large

Ex: 3 cars each driven w/ 4 ≠ brands of gasoline. Do we get the same mileage for each of the 4 brands?

Table 9.4-2 Gas mileage data

	Gasoline				
Car	1	2	3	4	$\bar{X}_{i.}$
1	26	28	31	31	29
2	24	25	28	27	26
3	25	25	28	26	26
$\bar{X}_{.j}$	25	26	29	28	27

① $H_B: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $\quad H_1$: Not H_B

② Reject H_B if $F > F_{0.01}(3, 6)$

$$\frac{\text{SS(B)}/(4-1)}{\text{SS(E)}/[(3-1)(4-1)]} \geq 9.78 = F_{0.01}(3, 6).$$

We have

$$\text{SS(B)} = 3[(25-27)^2 + (26-27)^2 + (29-27)^2 + (28-27)^2] = 30;$$

$$\text{SS(E)} = (26 - 29 - 25 + 27)^2 + (24 - 26 - 25 + 27)^2 + \cdots + (26 - 26 - 28 + 27)^2 = 4.$$

Hence, the computed F is

$$\frac{30/3}{4/6} = 15 > 9.78, \Rightarrow \text{Reject } H_B \rightarrow \text{gasolines are } \neq$$

Table 9.4-3 ANOVA table for gas mileage data

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F	p-value
Row (A)	24	2	12	18	0.003
Column (B)	30	3	10	15	0.003
Error	4	6	2/3		
Total	58	11			

- > 1 VALUE PER CELL
  REPLICATES

$X_{ijk} \sim N(\mu_{ij}, \sigma^2)$ independent rvs.

$i = 1, 2, \ldots$ a → row
$j = 1, 2, \ldots$ b → col
$K = 1, 2, \ldots$ c → replicates

$a \times b \times c$ = n comb.

Table 9.4-4 Two-way ANOVA table, c observations per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \frac{\text{SS(A)}}{a-1}$	$\frac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \frac{\text{SS(B)}}{b-1}$	$\frac{\text{MS(B)}}{\text{MS(E)}}$
Factor AB (interaction)	SS(AB)	$(a-1)(b-1)$	$\text{MS(AB)} = \frac{\text{SS(AB)}}{(a-1)(b-1)}$	$\frac{\text{MS(AB)}}{\text{MS(E)}}$
Error	SS(E)	$ab(c-1)$	$\text{MS(E)} = \frac{\text{SS(E)}}{ab(c-1)}$	
Total	SS(TO)	$abc-1$		

$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$
↓ ith row effect + ↓ jth col effect + interaction effect

$\sum_{i=1}^{a} \alpha_i = 0 = \sum_{j=1}^{b} \beta_j = 0 = \sum_{i=1}^{a} \gamma_{ij} = 0 = \sum_{j=1}^{b} \gamma_{ij} = 0$

H_0 $\begin{cases} H_A: \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0 \\ H_B: \beta_1 = \beta_2 = \ldots = \beta_b = 0 \\ H_C: \gamma_{ij} = 0 \text{ ? no interaction} \end{cases}$ => if true

$: \frac{SS(A)}{\sigma^2} \sim \chi^2_{(a-1)}$

$: \frac{SS(B)}{\sigma^2} \sim \chi^2_{(b-1)}$

$\frac{SS(E)}{\sigma^2} \sim \chi^2_{[ab(c-1)]}$

$: \frac{SS(C)}{\sigma^2} = \frac{\sum_{k=1}^{c} (X_{ijk} - \bar{X}_{ij\cdot})^2}{\sigma^2} \sim \chi^2_{(c-1)}$

$SS(TO) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (X_{ijk} - \bar{X}_{\cdots})^2 = SS(A) + SS(B) + SS(AB) + SS(E) \sim \chi^2_{(abc-1)}$

Test H_A => $F = \frac{MS(A)}{MS(E)}$ Reject H_A when $\geq F_\alpha[(a-1)(b-1), (abc-1)]$

Test H_B => $F = \frac{MS(B)}{MS(E)}$ Reject H_B when $\geq F_\alpha[(a-1)(b-1), (abc-1)]$

Test H_{AB} => $F = \frac{MS(AB)}{MS(E)}$ Reject H_{AB} when $\geq F_\alpha[(a-1)(b-1), (abc-1)]$

don't need to test

→ if there is interaction
⇓
there is difference among the means

ASSUMPTIONS of ANOVA:
- Observations are random samples from the population
- are drawn from Normally dist. populations
- Variance of all populations is equal.
- Numerator and denominator of F test are independent.