MATH 2411 CHEAT SHEET

by Frank

R Code Basics

```
> x = c(3,1,4,1,5,9,2,6)
> median = median(x)            > sanple_variance = var(x)        # use df = n - 1
> x_bar = mean(x)               > standard_deviation = sd(x)      # use df = n - 1
> n = length(x)                 > print(x)
```

Basics

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

Bayes' theorem: $P(B|A) = P(A|B)\dfrac{P(B)}{P(A)} = \dfrac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$

Expectation

$E(X) = \mu = \sum_i x_i p(x_i) = \int_{-\infty}^{+\infty} x p(x) dx$ $\quad E(X_1 + X_2) = E(X_1) + E(X_2)$ $\quad$ (X, Y not necessarily independent)

$E(aX + b) = aE(X) + b$ $\quad E(X_1 X_2) = E(X_1)E(X_2)$ $\quad$ (X, Y independent)

$E(g(X)) = \sum_i g(x)p(x) = \int_{-\infty}^{+\infty} g(x)p(x)dx$

Variation

$Var(X) = \sigma_X^2 = E((X-\mu)^2) = \sum_i (x_i - \mu)^2 p(x_i) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$ $\quad Var(X) = E(X^2) - (E(X))^2$

$Var(X \pm Y) = Var(X) + Var(Y)$ $\quad$ (X, Y independent) $\quad Var(aX + b) = a^2 Var(X)$

	$F(x)$	$f(x)$	
Discrete	c.d.f.	p.m.f.	
Continuous	c.d.f.	p.d.f.	Calculate c.d.f. first, then derive p.d.f.

Joint distribution

$p(x, y) = P(X = x, Y = y)$, $\sum_{x,y} p(x, y) = 1$ $\quad P(X \le a, Y \le b) = \int_{-\infty}^{a}\int_{-\infty}^{b} p(x, y)dxdy$, $\iint_{\mathbf{R}^2} p(x, y)dxdy = 1$

$p(x) = \sum_y p(x, y)$, $p(y) = \sum_x p(x, y)$ $\quad p(x) = \int_{-\infty}^{\infty} p(x, y)dy$, $p(y) = \int_{-\infty}^{\infty} p(x, y)dx$

Binomial distribution (discrete)

$X \sim B(n, p)$ $\quad P(X = x) = C_n^x p^x (1-p)^{n-x}$ $\quad E(X) = np, \quad Var(X) = np(1-p)$

```
> dbinom(x, size, prob)                      # returns f(x), i.e. P(X = x)
> pbinom(q, size, prob, lower.tail = TRUE)   # returns F(q), i.e. P(X ≤ q)
> qbinom(p, size, prob, lower.tail = TRUE)   # returns q where P(X ≤ q) = p, i.e. F⁻¹(p)
> rbinom(n, size, prob)                      # returns n samples from B(size, p)
```

Poisson distribution (discrete)

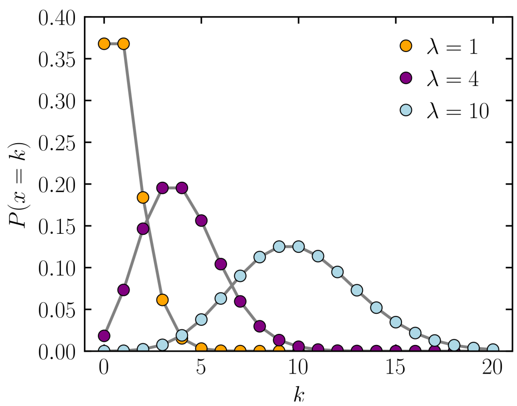
$X \sim \text{Pois}(n, p)$ $\quad P(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$ $\quad E(X) = \lambda, \quad Var(X) = \lambda$

```
> ppois(q, lambda, lower.tail = TRUE)
```

Normal distribution (continuous)

$X \sim \text{N}(\mu, \sigma^2)$ $\quad N(0,1) : f(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ $\quad E(X) = \mu, \quad Var(X) = \sigma^2$

$N(\mu, \sigma^2) : f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



$X \sim N(\mu, \sigma^2) \iff \dfrac{X-\mu}{\sigma} \sim N(0,1)$

$P(|X - \mu| < \sigma) \approx 0.683$

3σ-rule: $P(|X - \mu| < 2\sigma) \approx 0.954$, $\quad X \sim N(\mu, \sigma^2)$

$P(|X - \mu| < 3\sigma) \approx 0.997$

```
> qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)
```

Estimation

Estimator: distribution parameter, given random samples

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

Mean square error $MSE(\hat{\theta}, \theta) = E((\hat{\theta} - \theta)^2) = (\text{Bias}(\hat{\theta}, \theta))^2 + Var(\hat{\theta})$

Precision: $\dfrac{1}{\sigma_{\hat{\theta}}^2}$

Sample mean r.v. $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ $\qquad Var(\overline{X}) = \dfrac{\sigma_X^2}{n}$

$S_{n-1}^2 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ $\qquad E(S_{n-1}^2) = \sigma_X^2$ $\qquad Var(S_{n-1}^2) = \dfrac{1}{n}\left(\mu_4 - \dfrac{n-3}{n-2}(\sigma_X^2)^2\right)$, where $\mu_4 = E((X-\mu)^4)$

Maximum likelihood estimator $\hat{\theta}_{MLE}$: the $\hat{\theta}$ that maximizes $\prod_{i=1}^{n} p_\theta(x_i)$

Binomial $\hat{p} = \dfrac{\overline{X}}{m}$ $\qquad$ Poisson $\hat{\lambda} = \overline{X}$ $\qquad$ Normal $\hat{\mu} = \overline{X}$ $\quad \hat{\sigma}_{MLE}^2 = \dfrac{\sum_i (X_i - \overline{X})^2}{n}$ (biased)

Normal: $X \sim N(\mu, \sigma^2)$, we have $X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$, $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$

Poisson: $X \sim \text{Pois}(\lambda)$, we have $X_1 + X_2 + \cdots + X_n \sim \text{Pois}(n\lambda)$, $n\overline{X} \sim \text{Pois}(n\lambda)$

Binomial: $X \sim B(m, p)$, we have $X_1 + X_2 + \cdots + X_n \sim B(nm, p)$, $n\overline{X} \sim B(nm, p)$

Central limit theorem: $\lim_{n\to\infty} \dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$ for any distribution

Interval-valued estimation

Interval-valued estimation for $\overline{X}$ (when σ^2 is known)

CI for μ with $C = 1 - 2\alpha$: $\left[\overline{X} - z_\alpha \dfrac{\sigma}{\sqrt{n}}, \overline{X} + z_\alpha \dfrac{\sigma}{\sqrt{n}}\right]$, $\quad z_\alpha = \Phi^{-1}(1-\alpha)$ is called the critical value

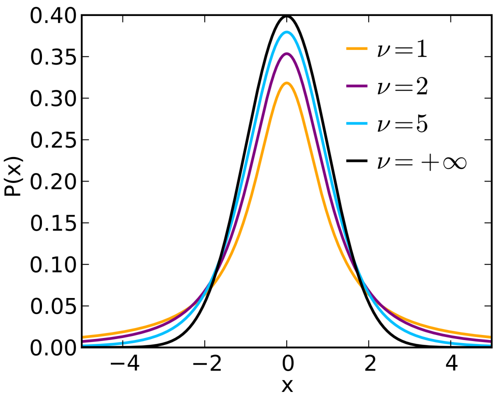
Interval-valued estimation for $\overline{X}$ (when σ^2 is unknown)

When $X \sim N(\mu, \sigma^2)$, we have $\dfrac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$

CI for μ with $C = 1 - 2\alpha$: $\left[\overline{X} - t_{n-1,\alpha}\dfrac{S_{n-1}}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha}\dfrac{S_{n-1}}{\sqrt{n}}\right]$

pdf of Student's t distribution: $f(t) = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \dfrac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

$\lim_{\nu\to\infty} t_\nu = N(0,1)$ $\qquad$ ν: degree of freedom



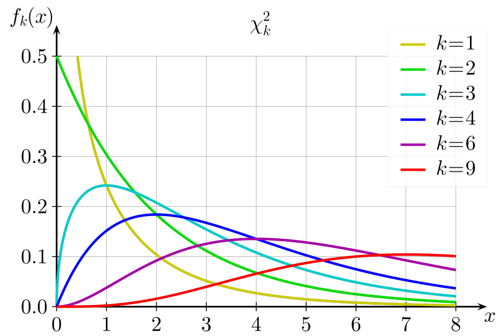

```
> qt(p, df, lower.tail = TRUE)          # df = n - 1
```

Interval-valued estimation for S_{n-1}^2:

When $X \sim N(\mu, \sigma^2)$, we have $\dfrac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi_{n-1}^2$

CI for σ^2 with $C = 1 - 2\alpha$: $\left[\dfrac{(n-1)S_{n-1}^2}{\chi_{n-1,\alpha}^2}, \dfrac{(n-1)S_{n-1}^2}{\chi_{n-1,1-\alpha}^2}\right]$

pdf of chi-sq distribution: $f(x;k) = \dfrac{x^{\frac{k}{2}+1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}$




```
> qchisq(p, df, lower.tail = TRUE)      # df = n - 1
```

Hypothesis testing

Null hypothesis H_0, an uninteresting explanation of the data

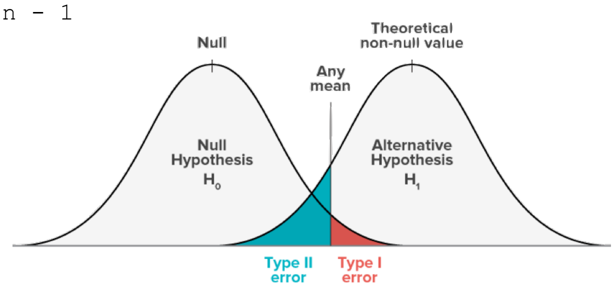
Alternative hypothesis H_1

Type I error $\alpha = P(H_0$ true but wrongly rejected)

Type II error $\beta = P(H_0$ false but wrongly retained)

smaller $\alpha \Rightarrow$ harder to reject H_0

power: $1 - \beta = P(H_0$ indeed rejected when it's false)




```
> power.t.test(n, delta, sd = 1, sig.level = 0.05, type = "one.sample", alternative =
"two.sided/one.sided")                  # delta = μ1 - μ0
                                        # then one.sided = greater
```

```
    # output:
      power, i.e. P(X > C) or P(X < C1) + P(X > C2) under H1
```

```
> power.t.test(delta, sd = 1, sig.level = 0.05, power, type = "two.sample", alternative =
"two.sided/one.sided")                      # delta = μ1 - μ0
                                            # then one.sided = greater
    # output:
      the min n that reaches the given power
```

Testing of μ when σ^2 is known

Idea: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

One-sided greater test

$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$

Rejection region: $\bar{X} > \mu_0 + z_\alpha \dfrac{\sigma}{\sqrt{n}}$

One-sided less test

$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

Rejection region: $\bar{X} < \mu_0 - z_\alpha \dfrac{\sigma}{\sqrt{n}}$

Two-sided test

$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$

Rejection region: $\bar{X} < \mu_0 - z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$ or $\bar{X} > \mu_0 + z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$

Simple test

$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{cases}$

CI for μ when σ^2 is known with $C = 1-\alpha$: $\left[\bar{X} - z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}\right]$, Reject H_0 if $\mu_0 \notin$ CI (two-sided test)

Testing of μ when σ^2 is unknown (t-test)

Idea: $\dfrac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$

t-value: $\dfrac{\bar{X}-\mu_0}{\frac{S_{n-1}}{\sqrt{n}}}$

Rejection region

One-sided greater test: $\bar{X} > \mu_0 + t_{n-1,\alpha} \dfrac{S_{n-1}}{\sqrt{n}}$, or equivalently $\dfrac{\bar{X}-\mu_0}{\frac{S_{n-1}}{\sqrt{n}}} > t_{n-1,\alpha}$

One-sided less test: $\bar{X} < \mu_0 - t_{n-1,\alpha} \dfrac{S_{n-1}}{\sqrt{n}}$, or equivalently $\dfrac{\bar{X}-\mu_0}{\frac{S_{n-1}}{\sqrt{n}}} < -t_{n-1,\alpha}$

Two-sided test: $\bar{X} < \mu_0 - t_{n-1,\frac{\alpha}{2}} \dfrac{S_{n-1}}{\sqrt{n}}$ or $\bar{X} > \mu_0 + t_{n-1,\frac{\alpha}{2}} \dfrac{S_{n-1}}{\sqrt{n}}$, or equivalently $\left|\dfrac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right| > t_{n-1,\frac{\alpha}{2}}$

CI for μ when σ^2 is unknown with $C = 1-\alpha$: $\left[\bar{X} - t_{n-1,\frac{\alpha}{2}} \dfrac{S_{n-1}}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}} \dfrac{S_{n-1}}{\sqrt{n}}\right]$, Reject H_0 if $\mu_0 \notin$ CI (two-sided test)

```
> t.test(x, alternative = "two.sided/less/greater", mu = 0, conf.level = 0.95)
    # output:
      t = t-value, df = n - 1, p-value
      alternative hypothesis: true mean is not equal to mu
      95 percent confidence interval:
      xxx    xxx                          # edge = (-)Inf for one-sided test
      mean of x
      xxx
```

p-value

p-value $= P\left(t \geq \dfrac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right) = \displaystyle\int_{\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}}^{+\infty} f(t)dt, \quad t \sim t_{n-1}$ (right test), $t_{n-1,\text{p-value}} =$ t-value

p-value $= P\left(t \leq \dfrac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right) = \displaystyle\int_{-\infty}^{\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}} f(t)dt, \quad t \sim t_{n-1}$ (left test)

p-value $= P\left(|t| \geq \left|\dfrac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right|\right) = \displaystyle\int_{-\infty}^{-\left|\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right|} f(t)dt + \int_{\left|\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right|}^{+\infty} f(t)dt$

$$= 2P\left(t \geq \left|\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right|\right) = 2\int_{\left|\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\right|}^{+\infty} f(t)dt, \quad t \sim t_{n-1} \text{ (two-sided test)}$$

Reject H_0 if p-value $\leq \alpha$

$$\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}} = \text{t-value} \quad > \quad t_{n-1,\alpha}$$

Relationships: $\quad t_{n-1,\cdot} \uparrow \downarrow P(t > \cdot)$ (rejection region)

$$\text{p-value} \quad \leq \quad \alpha$$

Testing of popular variance σ^2

Idea: $X \sim N(\mu, \sigma^2) \Rightarrow \dfrac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi_{n-1}^2$

$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$, Reject H_0 if $S_{n-1}^2 > \sigma_0^2 \dfrac{\chi_{n-1,\alpha}^2}{n-1}$, or equivalently, $\dfrac{(n-1)S_{n-1}^2}{\sigma_0^2} > \chi_{n-1,\alpha}^2$

$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$, Reject H_0 if $S_{n-1}^2 < \sigma_0^2 \dfrac{\chi_{n-1,1-\alpha}^2}{n-1}$, or equivalently, $\dfrac{(n-1)S_{n-1}^2}{\sigma_0^2} < \chi_{n-1,1-\alpha}^2$

$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$, Reject H_0 if $S_{n-1}^2 < \sigma_0^2 \dfrac{\chi_{n-1,1-\frac{\alpha}{2}}^2}{n-1}$ or $S_{n-1}^2 > \sigma_0^2 \dfrac{\chi_{n-1,\frac{\alpha}{2}}^2}{n-1}$,

CI for σ^2 with $C = 1-\alpha$: $\left[\dfrac{(n-1)S_{n-1}^2}{\chi_{n-1,\frac{\alpha}{2}}^2}, \dfrac{(n-1)S_{n-1}^2}{\chi_{n-1,1-\frac{\alpha}{2}}^2}\right]$, Reject H_0 if $\sigma_0^2 \notin$ CI (two-sided test)

or equivalently, $\dfrac{(n-1)S_{n-1}^2}{\sigma_0^2} < \chi_{n-1,1-\frac{\alpha}{2}}^2$ or $\dfrac{(n-1)S_{n-1}^2}{\sigma_0^2} > \chi_{n-1,\frac{\alpha}{2}}^2$

p-value $= P\left(U > \dfrac{(n-1)S_{n-1}^2}{\sigma_0^2}\right), \quad U \sim \chi_{n-1}^2$ (right test)

p-value $= 2 \cdot \min\left\{P\left(U < \dfrac{(n-1)S_{n-1}^2}{\sigma_0^2}\right), P\left(U > \dfrac{(n-1)S_{n-1}^2}{\sigma_0^2}\right)\right\}, \quad U \sim \chi_{n-1}^2$ (two-sided test)

$$\frac{(n-1)S_{n-1}^2}{\sigma_0^2} \quad > \quad \chi_{n-1,\alpha}^2$$

Relationships: $\quad \chi_{n-1,\cdot}^2 \uparrow \downarrow P(U > \cdot)$

$$\text{p-value} \quad \leq \quad \alpha$$

Testing of μ_X, μ_Y when σ_X^2, σ_Y^2 are known

Idea: $\begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim N(\mu_Y, \sigma_Y^2) \end{cases} \Rightarrow \begin{cases} \bar{X} \sim N\left(\mu_X, \dfrac{\sigma_X^2}{n}\right) \\ \bar{Y} \sim N\left(\mu_Y, \dfrac{\sigma_Y^2}{m}\right) \end{cases} \Rightarrow \bar{X}-\bar{Y} \sim N\left(\mu_X - \mu_Y, \dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}\right) \Rightarrow \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1)$

Two-sided test: $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$, Rejection region: $|\bar{X}-\bar{Y}| > z_{\frac{\alpha}{2}} \cdot \sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$, or equivalently, $\dfrac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} > z_{\frac{\alpha}{2}}$

CI for $\mu_X - \mu_Y$ with $C = 1-\alpha$: $\left[\bar{X}-\bar{Y}-z_{\frac{\alpha}{2}} \cdot \sqrt{\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{m}}, \ \ \bar{X}-\bar{Y}+z_{\frac{\alpha}{2}} \cdot \sqrt{\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{m}}\right]$, Reject H_0 if $0 \notin$ CI

Testing of μ_X, μ_Y when σ_X^2, σ_Y^2 are unknown but equal (two-sample t-test)

Pooled sample variance estimator: $S_p^2 = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_i - \bar{Y})^2}{n+m-2} = \dfrac{(n-1)S_{n-1,X}^2 + (m-1)S_{m-1,Y}^2}{n+m-2}$

Idea: $\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$

Two-sided test: $\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases}$ t-value: $\dfrac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$

Rejection region: $|\bar{X}-\bar{Y}| > t_{n+m-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}$, or equivalently, $\dfrac{|\bar{X}-\bar{Y}|}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} > t_{n+m-2,\frac{\alpha}{2}}$ (t-test)

One-sided greater test:
$\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X > \mu_Y \end{cases}$

Rejection region: $\bar{X}-\bar{Y} > t_{n+m-2,\alpha} \cdot S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}$

One-sided greater test:
$\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X < \mu_Y \end{cases}$

Rejection region: $\bar{X}-\bar{Y} < -t_{n+m-2,\alpha} \cdot S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}$

CI for $\mu_X-\mu_Y$ with $C = 1-\alpha$: $\left[\bar{X}-\bar{Y}-t_{n+m-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}, \bar{X}-\bar{Y}+t_{n+m-2,\frac{\alpha}{2}} \cdot S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}\right]$

Reject H_0 if $\mu_X-\mu_Y = 0 \notin$ CI

```
> t.test(x, y, alternative = "two.sided/less/greater", mu = 0, conf.level = 0.95,
var.equal = T)
    # output:
      t = t-value, df = n + m - 2, p-value
      alternative hypothesis: true difference in means is not equal to mu
      95 percent confidence interval:
      xxx    xxx
      mean of x     mean of y
      xxx           xxx
```

Testing of μ_X, μ_Y when σ_X^2, σ_Y^2 are unknown (Welch's t-test)

Idea: $\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_{n-1,X}^2}{n}+\frac{S_{m-1,Y}^2}{m}}} \approx t_\nu$, accurate enough when $n, m \geq 5$, where $\nu = \dfrac{\left(\dfrac{S_{n-1,X}^2}{n}+\dfrac{S_{m-1,Y}^2}{m}\right)^2}{\dfrac{1}{n-1}\left(\dfrac{S_{n-1,X}^2}{n}\right)^2+\dfrac{1}{m-1}\left(\dfrac{S_{m-1,Y}^2}{m}\right)^2}$

$\nu \in (\min\{n-1, m-1\}, n+m-2)$

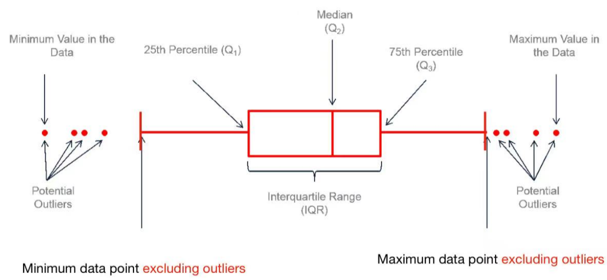
t-value: $\dfrac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{S_{n-1,X}^2}{n}+\frac{S_{m-1,Y}^2}{m}}}$ Welch's t-test: $\dfrac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{S_{n-1,X}^2}{n}+\frac{S_{m-1,Y}^2}{m}}} > t_{\nu,\frac{\alpha}{2}}$

$\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases}$, Rejection region: $|\bar{X}-\bar{Y}| > t_{\nu,\frac{\alpha}{2}} \cdot \sqrt{\dfrac{S_{n-1,X}^2}{n}+\dfrac{S_{m-1,Y}^2}{m}}$ or equivalently, $\dfrac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{S_{n-1,X}^2}{n}+\frac{S_{m-1,Y}^2}{m}}} > t_{\nu,\frac{\alpha}{2}}$

CI for $\mu_X-\mu_Y$ with $C = 1-\alpha$: $\left[\bar{X}-\bar{Y}-t_{\nu,\frac{\alpha}{2}} \cdot \sqrt{\dfrac{S_{n-1,X}^2}{n}+\dfrac{S_{m-1,Y}^2}{m}},\ \bar{X}-\bar{Y}+t_{\nu,\frac{\alpha}{2}} \cdot \sqrt{\dfrac{S_{n-1,X}^2}{n}+\dfrac{S_{m-1,Y}^2}{m}}\right]$

Reject H_0 if $\mu_X-\mu_Y = 0 \notin$ CI

```
> t.test(x, y, alternative = "two.sided/less/greater", mu = 0, conf.level = 0.95,
var.equal = F (default))
    # output:
      t = t-value, df = nu, p-value
      alternative hypothesis: true difference in means is not equal to mu
      95 percent confidence interval:
      xxx    xxx
      mean of x     mean of y
      xxx           xxx
```



Analysis of Variance

Boxplot
Points outside $[Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$ are potential outliers.

$\max\{Q_1 - 1.5\text{IQR}, \text{min value}\}$, $\min\{Q_3 + 1.5\text{IQR}, \text{max value}\}$
Factor: a categorical variable
Level: the possible value of a factor

Quantile: $Q_i = \begin{cases} x_k & , k \in \mathbb{Z} \\ \dfrac{x_{[k]} + x_{[k]+1}}{2} & , k \notin \mathbb{Z} \end{cases} \quad k = nq + 0.5, \quad i = \dfrac{q}{0.25}, \quad x_1 \le x_2 \le \cdots \le x_n$

Types of Variance

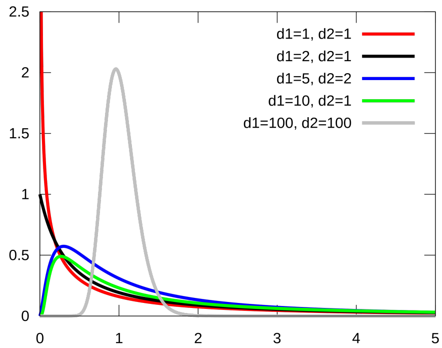
Anova model: Group i : $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n_i}$ are i.i.d. samples from $N(\mu_i, \sigma^2)$, $Y_{i,j} = \mu_i + \varepsilon_{i,j}$, i.i.d. $\varepsilon_{i,j} \sim N(0, \sigma^2)$

Total variance (sum of square total): $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y})^2$, $\text{df} = n_1 + \cdots + n_k - 1 = n - 1$, $n = \sum_{i=1}^{k} n_i$

Between variance (sum of square treatment) $SS_{\text{Treat}} = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2$, $\text{df} = k - 1$

Within variance (sum of square error): $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2$, $\text{df} = (n_1 - 1) + \cdots + (n_k - 1) = n - k$

Total variance = Between variance + Within variance, $SST = SS_{\text{Treat}} + SSE$



Mean square total: $MST = \dfrac{SST}{n-1}$

Mean square treatment: $MS_{\text{Treat}} = \dfrac{SS_{\text{Treat}}}{k-1}$

Mean square error: $MSE = \dfrac{SSE}{n-k} = \hat{\sigma}^2$

F statistics

F statistics: $F = \dfrac{\frac{SS_{\text{Treat}}}{k-1}}{\frac{SSE}{n-k}} = \dfrac{MS_{\text{Treat}}}{MSE}$

When σ^2 is unknown but uniform

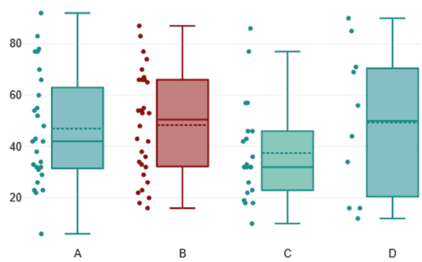
$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \text{Some } \mu_i \text{ are different} \end{cases}$, $Y_{i,j} = \mu_i + \varepsilon_{i,j}$, $\varepsilon_{i,j} \sim N(0, \sigma^2)$, σ^2 is unknown but uniform, $\varepsilon_{i,j}$ are i.i.d.

For Anova, we should reject H_0 if F is large.

F distribution: Under H_0, $F = \dfrac{MS_{\text{Treat}}}{MSE} \sim F(k-1, n-k)$

Reject H_0 if $F > F_{k-1, n-k, \alpha}$

When $k = 2$, two-sample two-sided t-test and Anova will give the same p-value, and $F = t^2$.




```
> qf(p, df1, df2, lower.tail = TRUE)     # df1 = k-1, df2 = n-k

> data = read.csv("Data_name.csv")
> na.omit()                             # omit the N/A entries
> alldata = c(data$X1,data$X2,data$X3,data$X4,data$X5)
> factor = c(rep("X1", n1), rep("X2", n2), rep("X3", n3))
> data = data.frame(Y = alldata, X = as.factor(factor))
> boxplot(Y ~ X, data = data)
> aov_result = aov(Y ~ X, data = data)
> summary(aov_result)                   # assuming equal variance
    # Output:
                    Df      Sum Sq        Mean Sq       F value       Pr(>F)
       x            k - 1   SSTreat       MSTreat       xxx           p-value ***
       Residuals    n - k   SSE           MSE
       ---
       Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When σ^2 is unknown and not uniform (Welch's Anova)

$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \text{Some } \mu_i \text{ are different} \end{cases}$, $Y_{i,j} = \mu_i + \varepsilon_{i,j}$, $\varepsilon_{i,j} \sim N(0, \sigma_i^2)$, σ_i^2 is unknown, $\varepsilon_{i,j}$ are i.i.d.

Welch's Anova: $F_W \stackrel{\rightarrow}{\sim} F(k-1, \frac{1}{\Lambda})$, hence we reject H_0 if $F_W > F_{k-1, \frac{1}{\Lambda}, \alpha}$

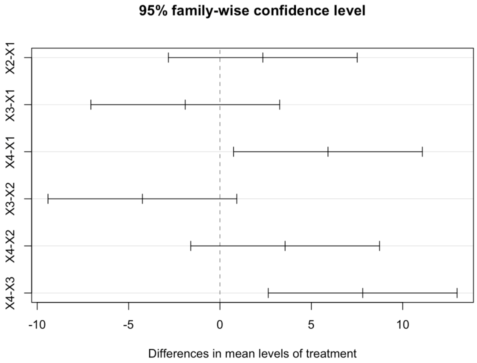
When $k = 2$, Welch's t-test and Welch's Anova will give the same p-value, and $F_W = t^2$.

```
> oneway.test(Y ~ X, data = data)         # not assuming equal variance
    # Output:
       data:  Y and X
       F = xxx, num df = k - 1, denom df = xxx, p-value = xxx
```

Tukey's Honestly Significant Difference (HSD)

$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$, $\forall i, j$, CI with $C = 1 - \alpha$: $\left[\bar{X}_i - \bar{X}_j - q_{k,n-k,\frac{\alpha}{2}} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \bar{X}_i - \bar{X}_j + q_{k,n-k,\frac{\alpha}{2}} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right]$, $s = \sqrt{MSE}$

Reject H_0 if $0 \notin$ CI




```
> TukeyHSD(aov_result)

> plot(TukeyHSD(aov_result))
```

Linear Regression

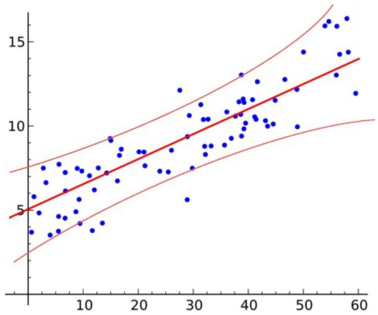
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, i.i.d. $\varepsilon_i \sim N(0, \sigma^2)$

Y: Dependent variable, response, regressand

x: Independent variable, explanatory variable, regressor (not random in this course)

$\hat{\beta}_0, \hat{\beta}_1$ minimize $\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2$

Notation: $\begin{cases} S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1) S_{n-1,x} \\ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (n-1) S_{n-1,Y} \end{cases}$



$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \end{cases}$ $\hat{\beta}_0, \hat{\beta}_1$ are unbiased

$Var(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$, $Var(\hat{\beta}_0) = \dfrac{\sigma^2 \overline{x^2}}{S_{xx}} := \dfrac{\frac{\sigma^2}{n} \sum_{i=1}^{n} x_i^2}{S_{xx}}$

Idea: $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$, $\hat{\beta}_0 \sim N\left(\beta_0, \dfrac{\sigma^2 \overline{x^2}}{S_{xx}}\right)$, or equivalently, $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$, $\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2 \overline{x^2}}{S_{xx}}}} \sim N(0,1)$

Two-sided test: $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$

Reject H_0 if $|\hat{\beta}_1| > z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{S_{xx}}}$, or equivalently, $\dfrac{|\hat{\beta}_1|}{\frac{\sigma}{\sqrt{S_{xx}}}} > z_{\frac{\alpha}{2}}$

CI for $\hat{\beta}_1$ with $C = 1 - \alpha$ when $\sigma^2 = Var(\varepsilon_i)$ is known: $\left[\hat{\beta}_1 - z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{S_{xx}}}, \hat{\beta}_1 + z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{S_{xx}}} \right]$

CI for $\hat{\beta}_0$ with $C = 1 - \alpha$ when $\sigma^2 = Var(\varepsilon_i)$ is known: $\left[\hat{\beta}_0 - z_{\frac{\alpha}{2}} \sqrt{\dfrac{\sigma^2 \overline{x^2}}{S_{xx}}}, \hat{\beta}_0 + z_{\frac{\alpha}{2}} \sqrt{\dfrac{\sigma^2 \overline{x^2}}{S_{xx}}} \right]$

Risidual: $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Mean Squared Error (MSE): $S^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2} = \dfrac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$ is unbiased for estimating $Var(\varepsilon)$, but $\sigma^2_{MLE} = \dfrac{\sum_{i=1}^{n} e_i^2}{n}$ is.

Idea: $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{S_{xx}}}} \sim t_{n-2}$, $\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{S^2 \overline{x^2}}{S_{xx}}}} \sim t_{n-2}$

$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$, Reject H_0 if $|\hat{\beta}_1| > t_{n-2,\frac{\alpha}{2}} \dfrac{S}{\sqrt{S_{xx}}}$, or equivalently, $\dfrac{|\hat{\beta}_1|}{\frac{S}{\sqrt{S_{xx}}}} > t_{n-2,\frac{\alpha}{2}}$

β_1 describes the strength of the linear relation.

CI for $\hat{\beta}_1$ with $C = 1 - \alpha$ when $\sigma^2 = Var(\varepsilon_i)$ is unknown: $\left[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \dfrac{S}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \dfrac{S}{\sqrt{S_{xx}}} \right]$

CI for $\hat{\beta}_0$ with $C = 1 - \alpha$ when $\sigma^2 = Var(\varepsilon_i)$ is unknown: $\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \cdot S \sqrt{\dfrac{\overline{x^2}}{S_{xx}}}, \hat{\beta}_0 + t_{n-2,\frac{\alpha}{2}} \cdot S \sqrt{\dfrac{\overline{x^2}}{S_{xx}}} \right]$

```
> plot(x, y, xlab = "X", ylab = "Y")           # Scatter plot
> lm(formula = Y ~ X, data = alldata)          # Linear model
    # Output:
      (Intercept)    X
```

```
        β0              β1
> abline(a = β0, b = β1, col = "red")          # Draw a line
> summary(lm(formula = Y ~ X, data = alldata))
    # Output:
        Residuals:
        Min     1Q      Median3Q      Max
        xxx     xxx     xxx     xxx     xxx

        Coefficients:
                        Estimate      Std. Error    t value       Pr(>|t|)
        (Intercept)     β0            S√x̄²/√Sxx      xxx           p-value *
        X               β1            S/√Sxx        xxx           p-value ***
        ---
        Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
        Residual standard error: √MSE on n - 2 degrees of freedom
        Multiple R-squared: R^2                 Adjusted R-squared: xxx
        F-statistic: xxx on x and n - 2 DF      p-value: xxx
```

Prediction

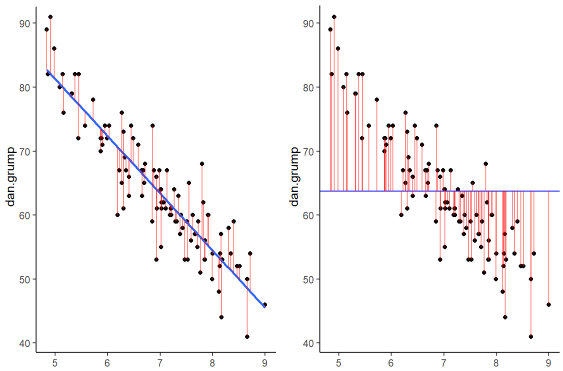
Idea: $\hat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ is unbiased, and $\dfrac{\hat{Y}_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})}{S\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

CI for $\hat{Y}_{\text{new}}$ with $C = 1 - \alpha$: $\left[\hat{Y}_{\text{new}} - t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \hat{Y}_{\text{new}} + t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}\right]$

Prediction Interval for $\hat{Y}_{\text{new}}$ with $C = 1 - \alpha$:

$$\left[\hat{Y}_{\text{new}} - t_{n-2\frac{\alpha}{2}} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \hat{Y}_{\text{new}} + t_{n-2\frac{\alpha}{2}} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}\right]$$

```
> Y_hat = lm(formula = Y ~ X, data = alldata)
> predict(Y_hat, data.frame(X = x_new))
    # Output
        1
        Ŷ_new
> predict(Y_hat, data.frame(X = x_new), interval = "confidence", level = 0.95)
    # Output
                fit     lwr     upr
        1       Ŷ_new   xxx     xxx
> predict(Y_hat, data.frame(X = x_new), interval = "prediction", level = 0.95)
    # Output
                fit     lwr     upr
        1       Ŷ_new   xxx     xxx
```



Decomposition of Variance

Total variance (SST): $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, df $= n - 1$

Regression variance (RSS): $RSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$, df $= 1$

Residual variance (SSE): $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, df $= n - 2$

$SST = RSS + SSE$

Coefficient of determination: $R^2 = \dfrac{RSS}{SST} = 1 - \dfrac{SSE}{SST} = \dfrac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \dfrac{S_{xy}^2}{S_{xx}S_{yy}} \in [0,1]$

Pearson's correlation coefficient: $r = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \in [-1,1]$