Statistics

Marie Biolkova

Useful Properties

Always:

$$\mathbb{E}(aX+b) = a\mathbb{E}(X)+b$$
$$\mathbb{E}(X+Y) = \mathbb{E}(X)+\mathbb{E}(Y)$$
$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

Only if *independent*:

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$$
$$\mathrm{Var}(X-Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$$
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i)$$

Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Unbiased and consistent estimator of μ.

Sample Variance

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

- Unbiased and consistent estimator of σ^2.
- Since $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ we have that $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ so we can estimate $\mathrm{Var}(\bar{X})$ by $\frac{S^2}{n}$.

Sample Covariance and Correlation

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \qquad \text{(covariance)}$$
$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \qquad \text{(correlation)}$$

Sample Covariance:

$$\begin{aligned} S_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)\right] \end{aligned}$$

- Unbiased and consistent estimator of $\mathrm{Cov}(X,Y)$
- S_{xx} and S_{yy} are the sample variances for X and Y, recall $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Sample Correlation:

$$R_{xy} = \frac{S_{xy}}{S_xS_y}$$

Maximum Likelihood Estimators (MLEs)

Assuming the data are independent, the *likelihood function* is

$$L(\theta; x_1,...,x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

The *log-likelihood* is therefore

$$l(\theta; x_1,...,x_n) = \sum_{i=1}^{n}\log f(x_i;\theta).$$

- $\hat{\sigma}^2$ is not the sample variance S^2.
- In general MLEs are biased estimators.
- Consistent estimators.

Invariance Property of MLEs: Let $\hat{\theta}$ be the MLE of θ and g be any function of θ. Then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Properties of the Sample Mean and Variance for the Normal Distribution

Let $X_1,..,X_n$ be independent $N(\mu,\sigma^2)$ rvs, then

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

Normal Distribution with Known Variance

Assume $X_1,...,X_n \sim N(\mu,\sigma^2)$ are independent rvs, σ^2 known. Recall $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ then the linear transform

$$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$$

is such that $Z \sim N(0,1)$.
The $(1-\alpha)\%$ *confidence interval* for μ is given by

$$\bar{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

- To calculate $z_{\alpha/2}$ in `R` use `qnorm(1-alpha/2, 0,1)`, e.g. `qnorm(0.975, 0,1)` for 95% CI.
- CI is larger for smaller sample size.
- Higher % confidence interval results in wider interval.

Normal Distribution with Unknown Variance

Assume $X_1,...,X_n \sim N(\mu,\sigma^2)$ are independent rvs, σ^2 unknown. Consider

$$T = \frac{\bar{X}-\mu}{\sqrt{S^2/n}}.$$

χ^2 Distribution

Let $Z_1,..,Z_n$ be independent $N(0,1)$ rvs and $X = \sum_{i=1}^{n} Z_i^2$. Then X has chi-squared distribution with n degrees of freedom, $X \sim \chi^2_n$.

- X is a continuous rv and $x \geq 0$.
- Let $Z \sim N(0,1)$ and $Y = Z^2$. Then $Y \sim \chi^2_1$.
- Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$, independently. Then $X+Y \sim \chi^2_{n+m}$.
- If $X \sim \chi^2_n$ then $\mathbb{E}(X) = n$ and $\mathrm{Var}(X) = 2n$.

t Distribution

Let X and Y be independent rvs such that $Z \sim N(0,1)$ and $Y \sim \chi^2_n$. Let $T = \frac{Z}{\sqrt{Y/n}}$, then T has a t-distribution with n degrees of freedom, i.e. $T \sim t_n$.

- T is a continuous rv, $t \in \mathbb{R}$.
- As $n \to \infty, t_n \to N(0,1)$.
- If $T \sim t_n$ the $\mathbb{E}(T) = 0$ and $\mathrm{Var}(T) = n/(n-1)$ for $n > 2$.
- Denote $t_{n;\alpha}$ the *upper α quantile*, i.e. $\mathbb{P}(T \geq t_{n;\alpha}) = \alpha$.
- Symmetrical about 0.

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

The $(1-\alpha)\%$ *confidence interval* for μ is

$$\bar{x} \pm t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}.$$

- To calculate $t_{n-1;\alpha/2}$ in `R` use `qt(1-alpha/2, n-1)`.
- The CI is larger when the variance is unknown.

Hypothesis Testing

- *Type I error*: Reject H_0 when it is in fact true.
- *Type II error*: Fail to reject H_0 when it is false.
- *Significance level α*: Probability that we reject H_0 when it is true, $\mathbb{P}(\text{Type I error}) = \alpha$.
- *Power β*: Probability that we reject H_0 when it is false, $\mathbb{P}(\text{Type II error}) = 1-\beta$.
- *Power function*: $\beta(\theta) = \mathbb{P}(\text{reject } H_0 : \theta = \theta_0 \text{ when the true value is } \theta)$.
- *Test statistic*: Function of the data chosen, is expected to take a different range of values when H_0 is true than when it is false.
- *Critical region C* The set of values of t that lead us to reject H_0.
- *p-value* is the probability of observing a result at least as extreme as t if H_0 is true.
  - p-value small ($< \alpha$): reject H_0.
  - p-value large ($\geq \alpha$): no evidence to reject H_0.

Increasing sample size means we are more likely to reject H_0 if it is false.

z-test

$X_1, ..., X_n$ independent $N(\mu, \sigma^2)$ rvs, σ^2 known.

1. $H_0: \mu = \mu_0 \quad$ vs $\quad H_1: \mu \neq \mu_0$
2. Test statistic: $T = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$, then under H_0, $T \sim N(0, 1)$.
3. Critical region: $|T| = \left|\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}$.
4. p-value: $\mathbb{P}(|T| \geq t_0) = 2\mathbb{P}(T \geq t_0) = 2\mathbb{P}(T \leq -t_0)$[1].

One Sample t-test

$X_1, ..., X_n$ independent $N(\mu, \sigma^2)$ rvs, σ^2 unknown.

1. $H_0: \mu = \mu_0 \quad$ vs $\quad H_1: \mu \neq \mu_0$
2. Test statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$, then under H_0, $T \sim t_{n-1}$.
3. Critical region: reject H_0 if $|T| \geq t_0 = t_{n-1;\alpha/2}$.
4. p-value: $\mathbb{P}(|T| \geq t_0) = 2\mathbb{P}(T \geq t_0) = 2\mathbb{P}(T \leq -t_0)$

Paired t-test

Paired data $(X_1, Y_1), ...(X_n, Y_n)$ where the two measurements are dependent. Consider the difference such that $D_i = Y_i - X_i$ for $i = 1, ..., n$.

Assume $D_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ - observed differences are independent of each other and observations are from normal distribution with mean μ and unknown variance σ^2.

Reduces to a one-sample t-test.

1. $H_0: \mu = 0 \quad$ vs $\quad H_1: \mu \neq 0$
2. Test statistic: $T = \dfrac{\bar{D}}{S/\sqrt{n}}$, then under H_0, $T \sim t_{n-1}$.

Two Sample t-test

Suppose we have two sets of independent rvs $X_1, ..., X_n$ and $Y_1, ..., Y_m$ such that $X_i \sim N(\mu_X, \sigma^2), Y_i \sim N(\mu_y, \sigma^2)$.

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1}$$

Pooled sample variance:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

1. $H_0: \mu_X = \mu_Y \quad$ vs $\quad H_1: \mu_X \neq \mu_Y$
2. Test statistic: $T = \dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$, then under H_0, $T \sim t_{m+n-2}$.
3. Critical region: reject H_0 if $|T| \geq t_0 = t_{m+n-2;\alpha/2}$.
4. p-value: $\mathbb{P}(|T| \geq t_0) = 2\mathbb{P}(T \geq t_0) = 2\mathbb{P}(T \leq -t_0)$

[1] t_0 is the upper quantile, $-t_0$ is the lower quantile.

F-test for Equality of Variance

Suppose we have two independent normal rvs $X_1, ..., X_n$ and $Y_1, ..., Y_m$ with variances σ_X^2, σ_Y^2.

1. $H_0: \sigma_X^2 = \sigma_Y^2 \quad$ vs $\quad H_1: \sigma_X^2 \neq \sigma_Y^2$
2. Test statistic: $T = \dfrac{S_X^2}{S_Y^2}$, then under H_0, $F \sim F_{n-1,m-1}$.

F Distribution

$U \sim \chi^2_m, V \sim \chi^2_n$ independent rvs. Then $X = \dfrac{U/m}{V/n}$ has an F distribution with m, n degrees of freedom ($X \sim F_{m,n}$).

- $1/X \sim F_{n,m}$
- Upper α quantile $F_{m,n;\alpha}$ is such that $\mathbb{P}(X \geq F_{m,n;\alpha}) = \alpha$, lower quantile $F_{m,n;1-\alpha} = 1/F_{n,m;\alpha}$.
- `pf` and `qf` commands in `R`

One-sided Tests

$H_0: \theta = \theta_0 \quad$ vs $\quad H_1: \theta > \theta_0$
$H_0: \theta = \theta_0 \quad$ vs $\quad H_1: \theta < \theta_0$

Linear Regression

$$\mathbb{E}(Y) = \alpha + \beta x$$

Least-Squares Estimation

Want to find $\hat{\alpha}, \hat{\beta}$ that minimise the sum of squares

$$S(\alpha, \beta) = \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 = \sum_{i=1}^{n}\epsilon_i^2.$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{S_{XY}}{S_{XX}}$$

- Requires no assumptions about the distribution.
- $\hat{\alpha}, \hat{\beta}$ are rvs, unbiased and consistent estimators of α, β.

Simple Linear Regression

Assume $Y_1, ..., Y_n$ are independent, normally distributed rvs with common variance, and have a mean that is a linear function of the explanatory variable, i.e
$Y_i \overset{\text{iid}}{\sim} N(\alpha + \beta x_i, \sigma^2) \quad i = 1, ..., n.$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right) \qquad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{XX}}\right)$$

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \text{ (fitted value)}$$

- S^2 is an unbiased estimator of σ^2 with $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$.
- S^2 is independent of $\hat{\alpha}, \hat{\beta}$ (but $\hat{\alpha}, \hat{\beta}$ are not independent!)
- *Standard errors*: s.e.$(\hat{\alpha}) = \sqrt{\text{Var}(\hat{\alpha})}$, s.e.$(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})}$
- Confidence intervals:

$$\hat{\alpha} \pm t_{n-2;0.025} \times s.e.(\hat{\alpha})$$
$$\hat{\beta} \pm t_{n-2;0.025} \times s.e.(\hat{\beta}).$$

Regression using `R`

Command `lm(y~x)`, and `lm(y~x - 1)` for regression through the origin.

Confidence Interval for $\mathbb{E}(Y_0)$

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

- Interval for the predicted expectation - they reflect uncertainty in our estimates of average observation.

Prediction Interval for Y_0

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

- Prediction for a single observation as a function of the explanatory variable - we would expect 95% of observations to lie within this interval.
- Prediction intervals for Y_0 are wider than confidence intervals for $\mathbb{E}(Y_0)$ as they take into account uncertainty relating to the expected value and individual variability.
- Confidence and prediction intervals become wider as x_0 moves away from $\bar{x}$.
- Do not extrapolate beyond the range of data as this is might be very inaccurate.

Multiple Regression

Assume $Y_i \sim N(\alpha + \beta_1 x_{1i} + .. + \beta_k x_{ki}, \sigma^2)$ for $i = 1, ..., n$ with $Y_1, ..., Y_n$ independent, i.e. the observations are independent, normally distributed, have constant variance and the expectations are linearly related to explanatory variables.

$$\mathbb{E}(Y) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

The least-squares estimates $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_k$ are values that minimise

$$S(\alpha, \beta_1, ..., \beta_k) = \sum_{i=1}^{n}[y_i - (\alpha + \beta_{1i}x_i + ... + \beta_k x_{ki})]^2.$$

$$S^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}[Y_i - (\hat{\alpha} + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki})]^2$$

Confidence intervals:

$$\hat{\alpha} \pm t_{n-(k+1);0.025} \times s.e.(\hat{\alpha})$$
$$\hat{\beta}_j \pm t_{n-(k+1);0.025} \times s.e.(\hat{\beta}_j).$$

Residual sum of squares (rss): $\displaystyle\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$.

F-test for Model Comparison

Used to see whether or not the full model gives a significantly better fit than a submodel.
H_0 : the specified regression coefficients are zero
H_1 : there is no restriction on the regression coefficients

Analysis of Variance

One-way ANOVA

Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., k$ and $j = 1, ..., n_i$ independently for all Y_{ij}, i.e. the observations are from a normal distribution, independent, have a common variance and a mean only dependent on the group they are member of.

$$H_0 : \mu_1 = ... = \mu_k \quad \text{vs} \quad H_1 : \mu_1, ..., \mu_k \text{ are not all equal.}$$

Source	d.f.	SS	MS	F	p
Between	$k-1$	SS_B	MS_B	F	p
Error	$n-k$	SS_W	MS_W		
Total	$n-1$	SS_{Tot}			

In R, use `anova(lm( ))`. Need to express the explanatory variable using `as.factor`.

$$SS_{Tot} = SS_B + SS_W$$

Between groups mean square:

$$MS_B = \frac{SS_B}{k-1}$$

Within groups mean square:

$$MS_W = \frac{SS_W}{n-k} = s^2 \quad \text{(residual mean square)}$$

s is the *residual standard error*.
If H_0 is true then $F = \frac{MS_B}{MS_W} \sim F_{k-1,n-k}$.

Least Significant Differences (LSD)

$$t_{n-k;\alpha/2}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \text{ or } t_{n-k;\alpha/2}\sqrt{\frac{2s^2}{m}}$$

if the samples are of equal size.

Two-way ANOVA

Assume $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ where $\mu_{ij} = \alpha_i + \beta_j$, i.e. the observations are from a normal distribution, independent, have a common variance and a mean that is a function of effect of each group.
Consider b blocks, k treatments, $n = bk$.
Test 1 (block effect):

$$H_0 : \alpha_1 = ... = \alpha_b \text{ vs } H_1 : \alpha_1, ..., \alpha_b \text{ are not all equal}$$

Test 2 (treatment effect):

$$H_0 : \beta_1 = ... = \beta_k \text{ vs } H_1 : \beta_1, ..., \beta_k \text{ are not all equal}$$

Source	d.f.	SS	MS	F	p
Blocks	$b-1$	SS_B	MS_B	F_B	p_B
Treatment	$k-1$	SS_T	MS_T	F_T	p_T
Error	$(b-1)(k-1)$	SS_W	MS_W		
Total	$bk-1$	SS_{Tot}			

$$SS_{Tot} = SS_B + SS_T + SS_W$$

$$MS_B = \frac{SS_B}{b-1} \qquad MS_T = \frac{SS_T}{k-1}$$

$$MS_W = \frac{SS_W}{(b-1)(k-1)}$$

$$F_B = \frac{MS_B}{MS_W} \qquad F_T = \frac{MS_T}{MS_W}$$

LSD for two-way ANOVA

Block effect:

$$t_{(b-1)(k-1);\alpha/2}\sqrt{2\frac{s^2}{k}}$$

Treatment effect:

$$t_{(b-1)(k-1);\alpha/2}\sqrt{2\frac{s^2}{b}}$$

Two-way ANOVA with r replications

Source	d.f.	SS	MS	F	p
Blocks	$b-1$	SS_B	MS_B	F_B	p_B
Treatment	$k-1$	SS_T	MS_T	F_T	p_T
Error	$rbk-b-k+1$	SS_W	MS_W		
Total	$rbk-1$	SS_{Tot}			

LSD for two-way ANOVA with replications

Block effect:

$$t_{rbk-b-k+1;\alpha/2}\sqrt{2\frac{s^2}{rk}}$$

Treatment effect:

$$t_{rbk-b-k+1;\alpha/2}\sqrt{2\frac{s^2}{rb}}$$