Statistical Tests

When sampling from one normal distribution

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

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

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

When sampling from two normal distributions

$$\frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

$$\frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1}$$

Given values from binomial distributions

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim N(0,1)$$

Given values from poisson distributions

$$\frac{\bar{X}_1 - \bar{X}_2 - (\lambda_1 - \lambda_2)}{\sqrt{\bar{X}_1/n_1 + \bar{X}_2/n_2}} \sim N(0,1)$$

For a chi squared test

$$\sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2_{n-\text{num val to constrain}}$$

$$- \chi^2_{(r-1)(c-1)} \quad \text{[continuing to}$$

Point Estimation

Given a sample, we wish to estimate a population parameter.

Method of Moments

Equate population moments with sample moments and solve for the parameter. For cases with two parameters, you may take moments about the mean. For more use moments about zero.

Method of Maximum Likelyhood

Set L to be the probability of the sample given the parameter (θ). For exam,

$$L(\theta) = \prod_{i=0}^{n} f(x_i;\theta) \times P(X > y;\theta)^{m}$$ ← we have m samples where we know only that they're greater than y.

and solve $\partial/\partial\theta\, L(\theta) = 0$ or $\partial/\partial\theta \log L(\theta) = 0$ to find the maximum. [For multiple parameters, differentiate in each and solve the simultaneous eq^{n}s.] Remember to check the turning point is a maximum.

This estimator is unbiased, approximately normal, and has a variance given by the Cramér - Rao lower bound.

$$CRLB = \frac{1}{n\,E\{[\partial/\partial\theta \log f(X;\theta)]^2\}} = \frac{1}{E\{[\partial/\partial\theta \log L(\theta)]^2\}} = \frac{1}{-E[\partial^2/\partial\theta^2 \log L(\theta)]}$$

Bias and MSE

When estimating, define the bias as $E(g(X)) - \theta$ and the mean squared error as $MSE(g(X)) = E((g(X)-\theta)^2) = \text{Variance} + \text{bias}^2$

Baby Statistics

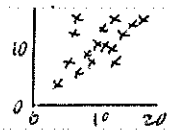
Bar charts, histograms, box plots and stem and leaf diagrams are time consuming to draw, so if you are asked to present data graphically, use a line plot or scatter diagrams.

1	5 6
2	1 3 3 4
2	5 5 5 7 8 8
3	0 3

Stem and leaf

Box plot

Line plot



Scatter diagram

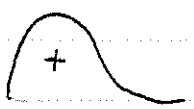
The range is the difference between the largest and smallest observations and the interquartile range is the difference between those at the quarter points.

The k^{th} moment about a point α is defined by $\frac{1}{n}\sum (x_i - \alpha)^k$

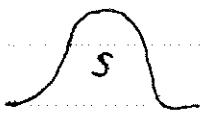
The mean is $\mu = \frac{1}{n}\sum x_i$ (the 1^{st} moment about zero), the median is the centre value and the mode is the most frequent value. We are only really interested in the mean.

The variance is $\sigma^2 = \frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}$ ($\frac{n}{n-1}$ times the 2^{nd} moment about μ)
The standard deviation is the positive square root of the variance.

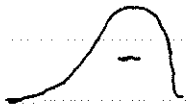
The population skew is hard to estimate from a sample. However if it has a tail in one direction, it will be skewed that way.



positively skewed



symmetrical



negatively skewed

Common Stats Process

1) List statistics gleaned from the data

$\sum x^2 = 0.2023$ $\quad \bar{x} = 0.11375$

$\sum x = 0.91$ $\quad \hat{\sigma}^2 = 0.01411$

$n = 8$

2) List the hypotheses to be tested

$H_0 : \mu = 0$

$H_1 : \mu > 0$

3) List the distribution, change until useful

$$\frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

$$\frac{\bar{x} - 0}{0.11880/\sqrt{8}} \sim t_7$$

4) List the probability, change until useful

$$P\left(\frac{\bar{x} - 0}{0.11880/\sqrt{8}} < 1.895\right) = 0.95$$

$$P(\bar{x} < 0.07960) = 0.95$$

5) List the conclusions drawn.

Since $0.11375 > 0.07960$, we have sufficient evidence at the 5% level to reject H_0, and conclude that $\mu > 0$.

Distributions

Single Parameter

The probability density function of X is $f_X(x) = P(X = x)$ if discrete and is defined by $\int_{x_1}^{x_2} f_X(x)\,dx = P(x_1 < X < x_2)$ if continuous.

The cumulative density function of X is $F_X(x) = P(X \leq x)$ if discrete and is defined by $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx = P(X \leq x)$ if continuous

Given a PDF, the k^{th} moment about a is $E[(X-a)^k] = \sum (x-a)^k f_X(x)$
$= \int (x-a)^k f_X(x)\,dx$

Thus we have mean $E(X)$, variance $E[(X - E(X))^2] = E(X^2) - E^2(X)$
and skew $E[(X - E(X))^3] = E(X^3) - 3E(X)E(X^2) + 2E^3(X)$

Finding the moments of linear combinations of distributions is easy.
$E(aX + bY + c) = aE(X) + bE(Y) + c$

Joint Distributions

The probability density function of X, Y is $f_{X,Y}(x,y) = P(X = x, Y = y)$ if discrete and defined by $\iint_A f_{X,Y}(x,y)\,dx\,dy = P(x_1 < X < x_2, y_1 < X < y_2)$ if continuous.

The cumulative density function of X, Y is $F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$ if discrete and defined by $\iint_A F_{X,Y}(x,y)\,dx\,dy = P(X \leq x, Y \leq y)$ if continuous.

To do anything useful, we need to convert this into a single parameter distribution.

- Ignore parameters $\quad f_X(x) = \int_y f_{X,Y}(x,y)\,dy$
- Conditional $\quad f_{X|Y=a}(x) = f_{X,Y}(x,a)$
- Convolution $\quad f_Z(z) = \int_x f_{X,Y}(x, z-x)\,dx \quad$ with $z = x+y$.

Variations on the last can give any function of the parameters.

Independance

If the values of two parameters in a joint distribution have no effect on each others distribution, we say they are independant. Note that three pairwise independant parameters may still be dependant on each other.

If X and Y are independant, then $f_X(x) = f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ and so $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. It immediately follows that

$$E(XY) = E(X)E(Y)$$
$$V(X+Y) = V(X) + V(Y) + 2E(XY) - 2E(X)E(Y) = V(X) + V(Y)$$

We define the covariance as $Cov(X,Y) = E(X,Y) - E(X)E(Y)$ and the correlation coefficient as $corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$

Analysis of Variance

Suppose we wish to test k samples (containing n total values) with normal dist^n and common variance, to see if they have the same mean. Let each value be denoted y_{ij} where i is the sample and j the value in the sample

	deg freedom	sums of squares	mean squares
between samples	$k-1$	$SS_B = \sum_i y_{i\cdot}^2/n_i - y_{\cdot\cdot}^2/n$	$SS_B/k-1$
residule	$n-k$	$SS_R = \sum_{ij} y_{ij}^2 - \sum_i y_{i\cdot}^2/n_i$	$SS_R/n-k$
total	$n-1$	$SS_T = \sum_{ij} y_{ij}^2 - y_{\cdot\cdot}^2/n$	

Reject H_0: samples have same mean if $\frac{SS_B}{k-1} / \frac{SS_R}{n-k} \sim F_{k-1,n-k}$ is too large.

Even if we reject H_0, we can use the student's t distribution to investigate results.
The confidence interval for the mean of a sample is $\bar{y}_{i\cdot} \pm t_{n-k}\,\hat{\sigma}\,(\frac{1}{n_i})^{1/2}$
The confidence interval for two samples to differ by is $\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{n-k}\,\hat{\sigma}\,(\frac{1}{n_i} + \frac{1}{n_j}$
By seeing if this interval includes zero, we can group treatments homogeneously.

$\bar{y}_{2\cdot} < \bar{y}_{3\cdot} < \bar{y}_{1\cdot} < \bar{y}_{5\cdot} < \bar{y}_{4\cdot}$

Generating Functions

Probability Generating Function $G_X(t) = E(t^X) = \sum_{x=0}^{\infty} t^x P(X=x)$

We can find moments by taylor expanding round $t=1$ to give

$$G_X(t) = 1 + (t-1)E(X) + \frac{(t-1)^2}{2!}E(X(X-1)) + \dots$$

Differentiating gives $G_X^{(n)}(1) = E(X(X-1)\dots(X-(n-1)))$

Note that $G_{a+bX}(t) = E(t^{a+bX}) = t^a E(t^{bX}) = t^a G_X(t^b)$

Moment Generating Function $M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$

We can find moments by expanding the exponential to give

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \dots$$

Differentiating gives $M_X^{(n)}(0) = E(X^n)$

Note that $M_{a+bX}(t) = E(e^{t(a+bX)}) = e^{at}E(e^{btX}) = e^{at}M_X(bt)$

Cumulant Generating Function $C_X(t) = \ln M_X(t) = \ln E(e^{tX})$

This generating function has the property that $C'_X(0) = E(X)$, $C''_X(0) = \mathrm{Var}(X)$ and $C'''_X(0) = \mathrm{Skew}(X)$.

Linear Regression

$$\bar{x} = \Sigma x_i / n$$
$$\bar{y} = \Sigma y_i / n$$
$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$
$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$$
$$S_{yy} = \sum y_i^2 - n\bar{y}^2$$

Suppose we wish to fit data to $Y_i = \alpha + \beta x_i + e_i$ $\quad E(e_i)=0,\ Var(e_i)=\sigma$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$\hat{\beta} = S_{xy}/S_{xx}$$
$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

Assuming e_i independant and identically normally distributed

$$E(Var(Y)) = \left(\frac{1}{n} + \frac{x_i - \bar{x}}{S_{xx}}\right)\sigma^2$$
$$est(Var(Y)) = \left(1 + \frac{1}{n} + \frac{x_i - \bar{x}}{S_{xx}}\right)\sigma^2$$
$$\hat{\beta} \sim N(\beta, \sigma^2/S_{xx})$$
$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Comparing expected with total variability gives the Coefficient of Determina
The square root of this is an estimator of the population correlation coefficien

$$R^2 = \left(\frac{S_{xy}^2}{S_{xx}}\right)/S_{yy} \qquad r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$H_0: \rho = 0 \quad \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$
$$H_0: \rho = \rho_0 \quad \tanh^{-1} r \sim N\left(\tanh^{-1}\rho, \frac{1}{n-3}\right)$$

Compound Distributions

Please note that $E(E(X|Y)) = \int \left(\int x f(x|y)\, dy \right) f(x)\, dx$

$= \iint x f(x,y)\, dx\, dy = E(X)$

and note that $V(X) = E(X^2) - E^2(X)$

$= E(E(X^2|Y)) - E^2(E(X|Y))$

$= E(E(X^2|Y)) - E(E^2(X|Y)) + E(E^2(X|Y)) - E^2(E(X|Y)$

$= E(V(X|Y)) - V(E(X|Y))$

Suppose we have a compound distribution S comprising of N distributions X.

$$E(S) = E(E(S|N))$$
$$= E(E(X_1 + \ldots + X_N | N))$$
$$= E(N E(X))$$
$$= E(N) E(X)$$

$$V(S) = E(V(S|N)) - V(E(S|N))$$
$$= E(V(X_1 + \ldots + X_N | N)) - V(E(X_1 + \ldots + X_N | N))$$
$$= E(N V(X)) - V(N E(X))$$
$$= E(N) V(X) - V(N) E^2(X)$$

$$M_S(t) = E(E(e^{tS}|N))$$
$$= E(E(e^{t(X_1 + \ldots + X_N)}|N))$$
$$= E(M_X(t)^N)$$
$$= E(e^{N \log M_X(t)})$$
$$= M_N(\log M_X(t))$$

$$G_S(t) = E(E(t^S|N))$$
$$= E(E(t^{X_1 + \ldots + X_N}|N))$$
$$= E(G_X(t)^N)$$
$$= G_N(G_X(t))$$