Statistics

Adrian Sai-wah Tam

March 31, 2006

1 Terminologies

- Descriptive statistics: Statistics that using pictures, etc to present data
- Inferential statistics: Statistics that conclude something form the data
- Measurements:
  - Ratio: Measurement with zero
  - Interval: Measurement without zero
  - Ordinal: Ordering
  - Nominal: Category identification
- Population: Everything in your interest
- Sample: The items that you examined/measured
- Sample size (n): The size of the sample
- Data (x_i): The information (value) obtained from sample
- Random variables (X): The name of the information
- r-th moment of X: $\overline{x^r} = E[X^r] = \frac{1}{n}\sum_{i=1}^{n} x_i^r$
  - a.k.a. r-th moment about zero
- r-th central moment of X: $m_r = E[(X-\bar{x})^r] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$
  - a.k.a. r-th moment about the mean $\bar{x}$
  - Dimensionless moment: $\alpha_r = m_r/\sqrt{m_2^r}$
- Skewness
  - Pearson's first coefficients of skewness: Skewness $= \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$
  - Pearson's second coefficients of skewness: Skewness $= \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$
  - Skewness is positive if the distribution is skewed to the right, i.e. having a longer tail to the right than to the left
    - known as right-skewed
    - the converse: left-skewed, with negative skewness
  - Moment coefficient of skewness: $\alpha_3 = m_3/\sqrt{m_2^3}$
- Kurtosis:
  - The degree of peakedness of a distribution, relative to normal distribution
  - *leptokurtic*: having a relatively high peak
  - *platykurtic*: flat-topped
  - *mesokurtic*: just like normal distribution, nor very peaked or very flat-topped
  - Moment coefficient of kurtosis: $\alpha_4 = m_4/m_2^2$

Sample vs Population

- Sample mean ($\bar{x}$): Mean obtained from samples, a known fact after survey. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Population mean (μ_X): Mean obtained from examination of the population, usually unknown but interested to know.
- Population variance: $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)^2$
  - Of finite population N
  - Also known as the second central moment of X
- Sample variance: $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - As an estimate of true population variance σ_X^2
  - Also known as the unbiased estimator and $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is called the biased estimator
    * As $\bar{X}$ is used as the mean instead of μ, the biased estimator is underestimating the true variance because there are raw counts of repeated elements
    * Proof of "unbiased": $E[s_X^2] = E[(x-\mu_X)^2]$

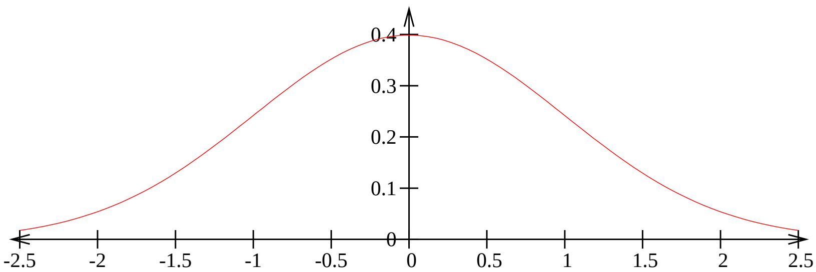
	Sample	Population
Size	n	N
Mean	$\bar{x} = \frac{1}{n}\sum x$	$\mu = \frac{1}{N}\sum x$
Variance	$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$	$\sigma^2 = \frac{1}{N}\sum(x-\mu)^2$
Standard Deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
Coefficient of Variation	$CV = s/\bar{x}$	$CV = \mu/\sigma$
z-Score	$z = (x-\bar{x})/s$	$z = (x-\mu)/\sigma$

2 Identities

$$\begin{aligned}
\mathrm{var}[X] &= E[(X-\mu_X)^2] \\
\mathrm{cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y \\
\mathrm{var}[aX+b] &= a^2\mathrm{var}[X] \\
\mathrm{var}[aX+bY] &= a^2\mathrm{var}[X] + b^2\mathrm{var}[Y] + 2ab\mathrm{cov}(X,Y) \\
E[aX+bY] &= aE[X] + bE[Y]
\end{aligned}$$

3 Normal Distribution

- Standard normal distribution: $f(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$
  - $\mu = 0, \sigma = 1$



- Commulative distribution of standard normal: $F(t) = \Pr[z \leq t] = \int_{-\infty}^{t} f(z)dz$
  - One-tail: $P(t) = \int_{-\infty}^{t} f(z)dz = 1 - \frac{1}{2}\text{erfc}\big(\frac{t}{\sqrt{2}}\big)$
  - Two-tail: $Q(t) = \int_{-t}^{t} f(z)dz = \text{erf}\big(\frac{t}{\sqrt{2}}\big)$
  - $\text{erf}(x) = 1 - \text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$
- General normal distribution: $N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Converting general normal distribution to standard normal distribution: $z = \frac{x-\mu}{\sigma}$
  - There are two points of inflexion: $\frac{d^2}{dx^2}N(x;\mu,\sigma) = 0$ at $x = \mu \pm \sigma$

3.1 Facts of Standard Distributions

- Binomial distribution
  - N trials, each with probability of success p, probability of failure $q = 1 - p$
  - Probability of X success out of N: $\Pr[X] = \binom{N}{X} p^X (1-p)^{N-X} = \binom{N}{X} p^X q^{N-X}$
- Poisson distribution
  - Probability of X arrivals with mean rate of λ: $\Pr[X] = \frac{\lambda^X e^{-\lambda}}{X!}$
- Cauchy distribution
  - Density function: $f(x) = \frac{a}{\pi(x^2+a^2)}$ with $a > 0$ and defined for $x \in (-\infty,\infty)$
- Gamma distribution
  - Density function: $f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$ with $\alpha,\beta > 0$ and defined for $x > 0$
  - $\Gamma(x) = \int_{0}^{\infty} t^{x-1}e^{-t}dt$ is the gamma function (for $x > 0$) with $\Gamma(n+1) = n!$ for integral n
- Beta distribution
  - Density function: $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ with $\alpha,\beta > 0$ and defined for $x \in (0,1)$
  - $B(m,n) = \int_{0}^{1} t^{m-1}(1-t)^{n-1}dt = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$ is the beta function (for $m,n > 0$)
- Geometric distribution
  - In a Bernoulli trial, the probability that the x-th trial is the first "success"
  - Mass function: $f(x) = p(1-p)^{x-1}$
- Pascal distribution
  - In a Bernoulli trial, the probability that the x-th trial can see the k-th "success"
  - Mass function: $f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$ for $x \geq k$
- Exponential distribution

- Density function: $f(x) = \lambda e^{-\lambda x}$

- Weibull distribution
  - Density function: $f(x) = abx^{b-1}e^{-ax^b}$

- Maxwell distribution
  - Modelling the magnitude of the speed of molecules in Brownian motion
  - Density function: $f(x) = \sqrt{2/\pi}\alpha^{3/2}x^2e^{-\alpha x^2/2}$

3.2 Approximation of Binomial Distribution

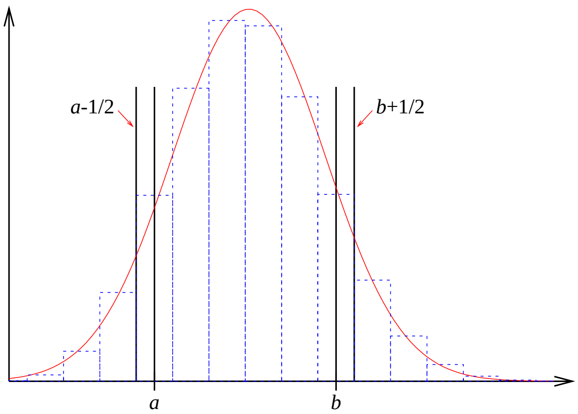
- n Bernoulli trials, each with probability of success p, the total number of success is a r.v. X
  - $X \sim \text{Binomial}(n, p)$
  - $\mu = np$
  - $\sigma^2 = np(1-p)$
  - Approximation by normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$:

$$\Pr[a \le X \le b] = \sum_{k=a}^{b} \binom{n}{k} p^k (1-p)^{n-k}$$

$$\begin{aligned}\Pr[a \le X \le b] &\approx \Pr\left[\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}} \le z \le \frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right] \\ &= \Pr\left[\frac{a-\frac{1}{2}-\mu}{\sigma} \le z \le \frac{b+\frac{1}{2}-\mu}{\sigma}\right]\end{aligned}$$

  - In general, using normal distribution to approximate a discrete distribution, we set

$$\Pr[a \le X \le b] = \Pr\left[\frac{a-\frac{1}{2}-\mu}{\sigma} \le z \le \frac{b+\frac{1}{2}-\mu}{\sigma}\right]$$



3.3 Approximation of Poisson Distribution

- In a system with exponential interarrival interval, the mean arrivals per unit time is λ
  - $X \sim \text{Poisson}(\lambda)$
  - $\mu = \sigma^2 = \lambda$

	Binomial	Poisson	Cauchy	Gamma	Beta	Geometric	Pascal	Exponential
Mean μ	Np	λ	0	$\alpha\beta$	$\frac{\alpha}{\alpha+\beta}$	$1/p$	k/p	$1/\lambda$
Variance σ^2	Npq	λ	∞	$\alpha\beta^2$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{1-p}{p^2}$	$\frac{k(1-p)}{p^2}$	$1/\lambda^2$
Moment coeff of skewness α_3	$\frac{q-p}{\sqrt{Npq}}$	$1/\sqrt{\lambda}$	undef					
Moment coeff of kurtosis α_4	$3+\frac{1-6pq}{Npq}$	$3+1/\lambda$	undef					
Mean deviation								
Moment generating func $M(t)$	$(q+pe^t)^N$	$e^{\lambda(\exp(t)-1)}$	undef	$(1-\beta t)^{-\alpha}$		$\frac{pe^t}{1-(1-p)e^t}$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^k$	$\frac{\alpha}{\alpha-t}$
Characteristic func $\phi(\omega)$	$(q+pe^{i\omega})^N$	$e^{\lambda(\exp(i\omega)-1)}$	$e^{-a\omega}$	$(1-\beta i\omega)^{-\alpha}$				

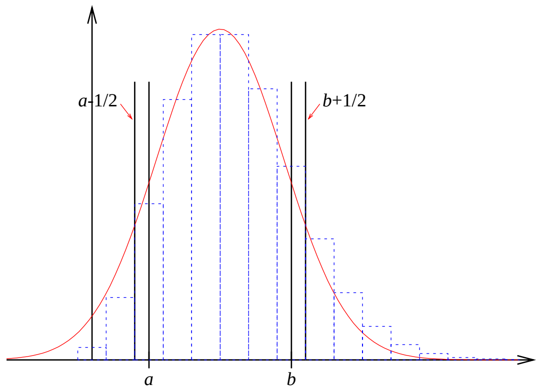
	Student's t	Chi-squared	F	Std Normal	Weibull	Maxwell
Mean μ	0	ν	$\frac{\nu_2}{\nu_2}-2$	0	$a^{-1/b}\Gamma\left(1+\frac{1}{b}\right)$	$2\sqrt{\frac{2}{\pi\alpha}}$
Variance σ^2	$\frac{\nu}{\nu-2}$	2ν	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2}$	1	$a^{-2/b}\left[\Gamma\left(1+\frac{2}{b}\right)-\Gamma^2\left(1+\frac{1}{b}\right)\right]$	$\frac{1}{\alpha}\left(3-\frac{8}{\pi}\right)$
Moment coeff of skewness α_3				0		
Moment coeff of kurtosis α_4				3		
Mean deviation				$\sigma\sqrt{2/\pi}$		
Moment generating func $M(t)$		$(1-2t)^{-\nu/2}$		$\exp(\mu t+\frac{1}{2}\sigma^2t^2)$		
Characteristic func $\phi(\omega)$		$(1-2i\omega)^{-\nu/2}$		$\exp(i\mu\omega-\frac{1}{2}\sigma^2\omega^2)$		

– Approximation by normal distribution with $\mu = \sigma^2 = \lambda$:

$$\Pr[a \le X \le b] = \sum_{k=a}^{b} \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\Pr[a \le X \le b] \approx \Pr\left[\frac{a - \frac{1}{2} - \lambda}{\sqrt{\lambda}} \le z \le \frac{b + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right]$$

$$= \Pr\left[\frac{a - \frac{1}{2} - \mu}{\sigma} \le z \le \frac{b + \frac{1}{2} - \mu}{\sigma}\right]$$



3.4 Central Limit Theorem

- For large sample size ($n \to \infty$) of a population, the z-score $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is close to normal
  – where μ is the population mean and σ is the population standard deviation
- For n samples x_i ($i = 1, \ldots, n$) from a population whose population mean is μ and variance σ^2, which are both finite, then

$$\lim_{n\to\infty} \Pr\left[a \le \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}} \le b\right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

4 Sampling Theory

- Sampling distributions of means:
  – Finite population of N, from which, taken n samples without replacement, then the expected sample mean and sample standard deviation would be

$$\bar{x} = \mu_X$$

$$s_X = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

  – If the population size is infinite, then $s_X = \sigma/\sqrt{n}$
  – For large samples ($n \ge 30$), by central limit theorem, the sampling distribution of means is approximately normal, so long as μ and σ are finite with $N > 2n$
- Sampling distribution of proportions
  – Infinite population, with n samples taken from, the probability of "success" in the population is p and that of "failure" is $q = 1 - p$
    * Binomial distribution

- The expected proportion of “success” in the sample and its standard deviation is

$$\bar{p} = \frac{E[x]}{n} = p$$
$$s_P = \frac{E[s_X]}{n} = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

- Sampling distribution of differences of means
  - Given two independent populations of infinite size, with n_1 samples drawn from the first population and n_2 samples drawn from the second population
  - The populations are having mean and standard deviations μ_1, σ_1 and μ_2, σ_2 respectively
  - Mean and standard deviations of the samples are $\bar{x}_1$, s_1 and $\bar{x}_2$, s_2 respectively
  - For $x = \bar{x}_1 - \bar{x}_2$, the expected value of x and its standard deviation are

$$\bar{x} = \mu_1 - \mu_2$$
$$s_x = \sqrt{s_1^2 + s_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

    * Also true for finite populations with samples taken with replcement
    * For finite population with samples taken without replacement, $s_x = \sqrt{s_1^2 + s_2^2} = \sqrt{\frac{\sigma_1^2}{n_1}\sqrt{\frac{N_1 - n_1}{N_1 - 1}} + \frac{\sigma_2^2}{n_2}\sqrt{\frac{N_2 - n_2}{N_2 - 1}}}$

- Sampling distribution of differences of proportions
  - Given two independent populations of infinite size, with n_1 samples drawn from the first population and n_2 samples drawn from the second population
  - Probability of “success” in the populations are p_1 and p_2 respectively
  - The proportion of success in the samples are $\bar{p}_1 = x_1/n_1$ and $\bar{p}_2 = x_2/n_2$ respectively
  - For $p = \bar{p}_1 - \bar{p}_2$, the expected value of p and its standard deviation are

$$\bar{p} = p_1 - p_2$$
$$s_p = \sqrt{s_1^2 + s_2^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Sampling distribution of sum of statistics
  - Given two samples, 1 and 2, the statistics are x_1 and x_2 with the respective standard deviation s_1 and s_2
  - Sum of statistics: $x_{1+2} = x_1 + x_2$
  - S.D. of statistics: $s_{1+2} = \sqrt{s_1^2 + s_2^2}$
- Standard errors: Standard deviation of a sampling distribution of a statistics is often called the *standard error*

5 Small Sampling Theory

- Small samples: $n < 30$

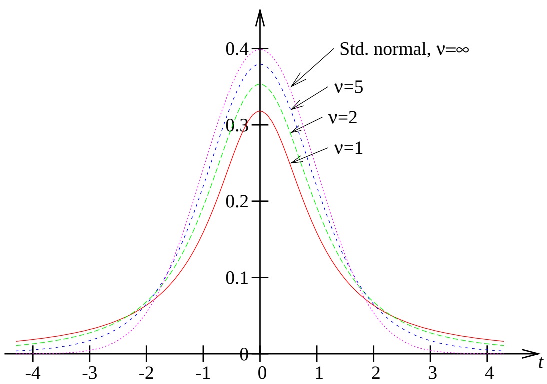
5.1 Student’s t-distribution

- Student’s t-distribution:

$$Y(t) = Y_0\left(1 + \frac{t^2}{n-1}\right)^{-n/2} = Y_0\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$
$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Statistics of sampling distribution	Expectation of estimate	Standard error of estimate	Remarks
Mean	$\bar{x} = \mu$	$s_x = \dfrac{\sigma}{\sqrt{n}}$	Close to normal when $n \geq 30$
Proportions	$x/n = p$	$s_p = \sqrt{\dfrac{p(1-p)}{n}}$	Close to normal when $n \geq 30$
Standard deviations	$s = \sigma$	$s_\sigma = \dfrac{\sigma}{\sqrt{2n}}$	For a normal population. Close to normal when $n \geq 100$
	$s \approx \sigma$	$s_\sigma = \sqrt{\dfrac{\mu_4 - \mu_2^2}{4n\mu_2}}$	For non-normal population. Here, μ_k is the k-th moments about the mean in the population
Medians	median	$s_{\text{med}} = \sigma\sqrt{\dfrac{\pi}{2n}} = \dfrac{1.2533\sigma}{\sqrt{n}}$	For a normal population. Close to normal when $n \geq 30$
1st or 3rd quartiles	$Q1$ or $Q3$	$s_{Q1} = s_{Q3} = \dfrac{1.3626\sigma}{\sqrt{n}}$	For a normal population. Close to normal when $n \geq 30$
Deciles	$D1$ or $D9$	$s_{D1} = s_{D9} = \dfrac{1.7094\sigma}{\sqrt{n}}$	For a normal population. Close to normal when $n \geq 30$
	$D2$ or $D8$	$s_{D2} = s_{D8} = \dfrac{1.4288\sigma}{\sqrt{n}}$	
	$D3$ or $D7$	$s_{D3} = s_{D7} = \dfrac{1.3180\sigma}{\sqrt{n}}$	
	$D4$ or $D6$	$s_{D4} = s_{D6} = \dfrac{1.2680\sigma}{\sqrt{n}}$	
Semi-inter-quartile range	$Q = Q3 - Q1$	$s_Q = \dfrac{0.7867\sigma}{\sqrt{n}}$	For a normal population. Close to normal when $n \geq 30$
Variance	$s^2 = \sigma^2$	$s_{\sigma^2} = \sigma\sqrt{\dfrac{2}{n}}$	For a normal population. Close to normal when $n \geq 100$
	$s^2 = \sigma^2\dfrac{n-1}{n}$	$s_{\sigma^2} = \sqrt{\dfrac{\mu_4 - \frac{n-3}{n-1}\mu_2^2}{n}}$	For a non-normal population.
Coefficient of variation	$s/\bar{x} = \sigma/\mu$	$s_{CV} = \dfrac{\sigma/\mu}{\sqrt{2n}}\sqrt{1+2(\sigma/\mu)^2}$	For a normal population. Close to normal when $n \geq 100$

- $\nu = n-1$ is the number of degrees of freedom
- Y_0 is a normalization constant for making $\int_{-\infty}^{\infty} Y(t)dt = 1$
- As $n \to \infty$, $Y(t)$ tends to standard normal distribution function



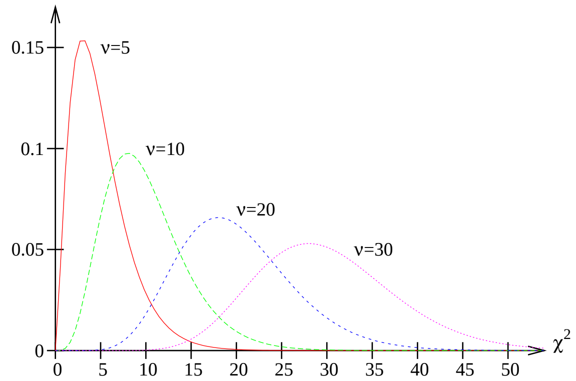
- Analogous to $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, we define $t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$
  where $s = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ is the unbiased estimate of the population standard deviation σ
  - For converting statistics $\bar{x}$ to fit into Student's t-distribution

5.2 Chi-square Distribution

- Chi-square distribution:

$$Y(\chi^2) = Y_0\left(\sqrt{\chi^2}\right)^{\nu-2} e^{-\chi^2/2} = Y_0 \chi^{\nu-2} e^{-\chi^2/2}$$
$$= \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \chi^{\nu-2} e^{-\chi^2/2}$$

  - where $\nu = n-1$ is the number of degrees of freedom
  - Y_0 is the normalization constant to make $\int_0^\infty Y(\chi^2)d\chi^2 = \int_{-\infty}^\infty \tilde{Y}(\chi)d\chi = 1$
  - Maximum value of $Y(\chi^2)$ attained when $\chi^2 = \nu - 2$



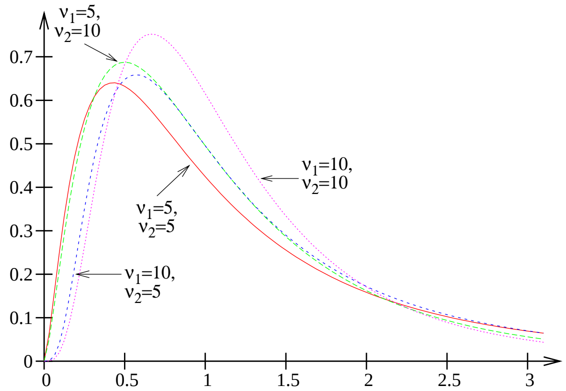
- Define $\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{\sum_i (x_i - \bar{x})^2}{\sigma^2}$

5.3 Fisher's F-distribution

- F-distribution

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$
$$Y(F) = Y_0 \frac{F^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}}$$
$$= \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{F^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}}$$

  - s_1 and s_2 are the unbiased estimation of standard deviations
  - $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are the degrees of freedom
  - Y_0 is the normalization constant to make $\int_0^\infty Y(F)dF = 1$



6 Correlation

- Pearson correlation factor:

$$r = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - E[X]^2}\sqrt{E[Y^2] - E[Y]^2}} = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\left[\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right]\left[\sum y^2 - \frac{1}{n}\left(\sum y\right)^2\right]}}$$

  - n samples taken from a population
  - r describes the correlation between statistic X and statistic Y of the sample
  - $-1 \leq r \leq 1$
    - $r = 0$: statistics X and Y are independent
    - $r \approx 0$: statistics X and Y do not have much correlation
    - $r \approx 1$: X and Y usually happen together
    - $r \approx -1$: X and Y usually contradicts
- Pearson correlation factor is suitable for linear relationships
  - Bimodal relation would give wrong conclusion

7 Confidence

- By CLT, everything can be converted to standard normal distribution
- In standard normal distribution, z lies in certain interval with some probability (for 30 samples or more)
  - If number of samples are small, or for higher accuracy (which depends on the number of samples), Student's t distribution should be used instead of standard normal distribution

Example: Mean

- Sampling $n \geq 30$ items from an infinite population
- Mean value of a property (e.g. weight) is $\bar{x}$ and the (sample) variance is s_X
- We say that,
  the mean value of the property (e.g. weight) of the population is $\bar{x} \pm z_{\alpha/2}\frac{s_X}{\sqrt{n}}$ with probability $1-\alpha$, where $z_{\alpha/2}$ satisfies $\frac{1}{\sqrt{2\pi}}\int_{-z_{\alpha/2}}^{z_{\alpha/2}} e^{-t^2/2}dt = 1-\alpha$
  - $z_{\alpha/2}$ is called the critical value
  - for smaller number of samples, $t_{\alpha/2}$ should be used instead of $z_{\alpha/2}$, and take $n-1$ degrees of freedom

Example: Bernoulli Trial

- For $n \geq 30$ trials, with success count of x
- Sample proportion of success is $p = x/n$, and the standard error is $\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{x/n(1-x/n)}{n}}$
- We say that,
  the success probability in the population is $p \pm z_{\alpha/2}(\mathrm{SE}) = \frac{x}{n} \pm z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}}$ with probability $1-\alpha$
  - $1-\alpha$ is called the *confidence level*
  - $p \pm z_{\alpha/2}(\mathrm{SE})$ is called the *confidence interval*

Example: Determining required sample size

- In a Bernoulli trial, sample for the rate of success
- Required accuracy: error to be within $\pm E$
- For a sample size of n, the number of success is x, then $\mathrm{SE} = \sqrt{\dfrac{x/n(1-x/n)}{n}}$ and the magnitude of error is $t_{\alpha/2}\sqrt{\dfrac{x/n(1-x/n)}{n}} \le E$

$$\begin{aligned}(t_{\alpha/2})^2\frac{x/n(1-x/n)}{n} &\le E^2 \\ \frac{1}{n} &\le \frac{E^2}{(t_{\alpha/2})^2 p(1-p)} \\ n &\ge \frac{(t_{\alpha/2})^2 p(1-p)}{E^2}\end{aligned}$$

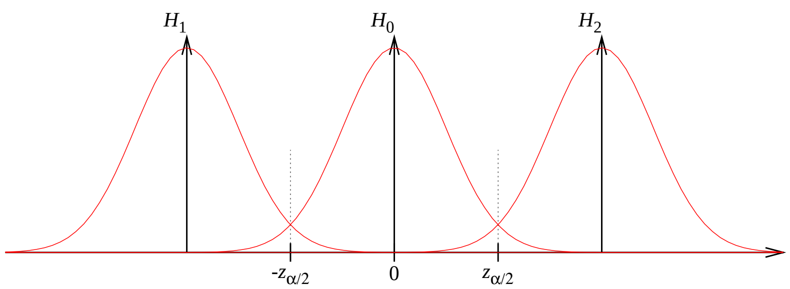
where $p = x/n$ is the raw measured success rate

8 Hypothesis Testing

- Given observation and hypothesis, test if the observation is a particular case of the hypothesis with certain probability

8.1 Procedure of hypothesis testing

1. Set up hypothesis
   - H_0: Null hypothesis, which usually represent the case that nothing special has been happended. Example: no effect on additional measure, no change ever happend
   - H_1: Alternative hypothesis, which represents the null hypothesis is false
   - H_2: Yet another alternative hypothesis (optional). If exists, H_1 and H_2 usually represents cases at the two "tails" respectively
2. Measurement, by survey or experiment to collect statistics
3. Verify the result of measurement, to see whether the probability that the occurance of the measurement result is acceptable within the hypothesis



Example: Is Diet Coke carcinogenic?

- According to survey, probability of having cancer by a normal people is μ
- Set:
  - H_0: Diet coke is not carcinogenic
  - H_1: Diet coke is carcinogenic
  - H_2: Diet coke is cancer-preventing (optional)
- Observing a group of n diet coke fans for a long time, and found that x of them turns out to have cancer

- Then the sample mean of cancer rate is $p = x/n$ and the standard error is $s_p = \sqrt{\frac{p(1-p)}{n}}$
- Assume that our predefined confidence level is $1-\alpha$, so we can find $z_{\alpha/n}$ (assume $n \geq 30$)
- Conclusion:
  - If $z = \frac{p-\mu}{s_p} < z_{\alpha/2}$, (or if H_2 defined, $|z| < z_{\alpha/2}$) then accept H_0
  - If $z = \frac{p-\mu}{s_p} \geq z_{\alpha/2}$, then accept H_1
  - (optional) If $z = \frac{p-\mu}{s_p} \leq -z_{\alpha/2}$, then accept H_2

Example: Are boys and girls having same weight?

- For a group of boys and a group of girls, which of size n_1 and n_2 respectively
- Mean weight and variance measured are x_1, σ_1^2 and x_2, σ_2^2 respectively
- Set:
  - H_0: Same weight, i.e. $x_1 - x_2 = 0$
  - H_1: not the same, i.e. $x_1 - x_2 \neq 0$
- Variable interested: $x_1 - x_2$
- Standard error: $\mathrm{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - Pooled variance estimate: $\sigma_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$
  - Alternative way of writing standard error: $\mathrm{SE} = \sqrt{\sigma_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
- Define: $z = \frac{(x_1-x_2)-0}{\mathrm{SE}} = \frac{x_1-x_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$
- Conclusion:
  - If $|z| < z_{\alpha/2}$, then accept H_0
  - If $|z| \geq z_{\alpha/2}$, then accept H_1

Example: How effective are the two drugs

- For two group of patients, which of size n_1 and n_2, and prescribed with drugs 1 and 2 respectively
- Number of patients cured in each group: x_1 and x_2,
- Set:
  - H_0: Two drugs are of same probability of effect, i.e. $p_1 - p_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} = 0$
  - H_1: Not the same
- Standard error: $\mathrm{SE} = \sqrt{(\mathrm{SE}_1)^2 + (\mathrm{SE}_2)^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- Define: $z = \frac{p_1 - p_2}{\mathrm{SE}}$
- Conclusion:
  - If $|z| < z_{\alpha/2}$, then accept H_0
  - If $|z| \geq z_{\alpha/2}$, then accept H_1
    - If $z \geq z_{\alpha/2}$, then drug 1 is more effective
    - If $-z \leq z_{\alpha/2}$, then drug 2 is more effective

Example: Hypothesis testing with small samples

- Generally the same procedure
- Instead of $z_{\alpha/2}$, take $t_{\alpha/2}$
  - Use Student's t distribution instead of standard normal distribution
  - Degree of freedom: $n-1$ for testing with single group of samples (one sample mean, $\bar{x}$, is used in the analysis)
  - Degree of freedom: n_1+n_2-2 for testing two two groups of samples (two sample means, $\bar{x}_1$ and $\bar{x}_2$, are used in the analysis)

8.2 Decision theory

- Confidence level: $1-\alpha$
- Power of test: $1-\beta$
- Type I error: Accepting H_1 when H_0 is true
  - Probability of type I error: α
- Type II error: Accepting H_0 when H_1 is true
  - Probability of type II error: β

Example: Error in digital communication

- Voltage transmitted: v_L volt for signal L and v_H volt for signal H, $v_L < v_H$
- White noise: $N(0,\sigma^2)$
- Received signal is decoded as signal L if the voltage received is $v < V$
  - If $v > V$, decoded as signal H
- Type 1 error: Erroneosly decoded signal L as H (one-tail)
  - Probability = α, where $z_\alpha = (V-v_L)/\sigma^2$
- Type 2 error: Erroreously decoded signal H as L (one-tail)
  - Probability = β, where $z_\beta = (v_H-V)/\sigma^2$
- Reducing Type 1 error: increase V
- Reducing Type 2 error: decrease V

9 Comparing Distribution

- Use Chi-square distribution

9.1 Goodness-of-fit test

- To determine whether the observed distribution model fits an expected model
- Define chi-square as: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i-E_i)^2}{E_i}$
  - k: number of category
  - O_i: Observed frequency of category i
  - E_i: Expected frequency of category i
- Compare using χ^2 distribution with $k-1$ degrees of freedom: the observation agrees with the expectation if $\chi^2 < \chi^2_{\alpha/2}$ for confidence level of $1-\alpha$

Example: Verifying a dice is unbiased

- Throwing a dice for n times, and record the number of times that different faces showed up
  - Let O_k be the number of times that face k showed up ($k = 1, \ldots, 6$)
- If it is a fair dice, the frequency for face k should be $E_k = n/6$
- Compute $\chi^2 = \sum_{k=1}^{6} \frac{(O_k - E_k)^2}{E_k} = \sum_{i=1}^{6} \frac{(O_k - n/6)^2}{n/6}$
- The critical value $\chi^2_{\alpha/2}$ should be found, with confidence level $1 - \alpha$ and $6 - 1 = 5$ degrees of freedom
- It is a fair dice if $\chi^2 < \chi^2_{\alpha/2}$

9.2 Independence test

- To determine whether different sets of samples are of the same distribution
- Observe r_k samples in set k
  - The frequency observed for category j in sample set k is O_{kj}
  - The sum of all frequencies in all the sample sets in category j is c_j
  - Total number of observations: $N = \sum_k r_k = \sum_j c_j$
  - If the sample sets are of the same distribution, then we expect the frequency of category j in sample set k to be $E_{kj} = \frac{r_k c_j}{N}$
- Define chi-square as: $\chi^2 = \sum_k \sum_j \frac{(O_{kj} - E_{kj})^2}{E_{kj}}$
  - Degrees of freedom: $(r-1)(c-1)$
    where r is the number of sample sets and c is the number of categories
- The different sample sets are in the same distribution if $\chi^2 < \chi^2_{\alpha/2}$ for confidence level of $1 - \alpha$

Example: Grading Standard in Courses

- To see whether different courses have the same standard in grading
- For the k courses, there are j grades awarded, and the number of students enrolled in course k is r_k
- Tabularize the number of students awarded to different grades in different courses:

	Grade 1	Grade 2	$\cdots$	Grade j	Total enrollment
Course 1	O_{11}	O_{12}	$\cdots$	O_{1j}	$r_1 = \sum_j O_{1j}$
Course 2	O_{21}	O_{22}	$\cdots$	O_{2j}	$r_2 = \sum_j O_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Course k	O_{k1}	O_{k2}	$\cdots$	O_{kj}	$r_k = \sum_j O_{kj}$
Total awardees	$c_1 = \sum_k O_{k1}$	$c_2 = \sum_k O_{k2}$	$\cdots$	$c_j = \sum_k O_{kj}$	$N = \sum_k \sum_j O_{kj}$

- Find $\chi^2 = \sum_k \sum_j \frac{(O_{kj} - E_{kj})^2}{E_{kj}}$ with $E_{kj} = \frac{r_k c_j}{N}$
- If $\chi^2 < \chi^2_{\alpha/2}$ for confidence level of $1 - \alpha$ and $(k-1)(j-1)$ degrees of freedom, then the courses are having the same grading standard, otherwise, there are some courses have biases.

10 ANOVA: Analysis of Variance

- Compare whether the means of two populations are equal: Two-way t-test (i.e. testing the difference of mean)
- Compare whether the means of several populations are equal: ANOVA
- Data:
  - k: Number of population
  - n_i: Sample size of population i
  - $\bar{x}_i$: Sample mean of population i
  - s_i^2: Sample variance of population i
  - μ_i: Population mean of population i
  - σ_i^2: Population variance of population i
  - $n = \sum_i n_i$: Total sample size
  - $\bar{x}$: Overall sample mean of the n samples
  - T_i: Sum of the n_i samples from population i
  - $\nu_1 = k - 1$: Degrees of freedom for treatments
  - $\nu_2 = n - k$: Degrees of freedom for error
  - $n - 1$: Degrees of freedom for total
- Data to compute:
  - Treatment sum of squares: $\text{SSTR} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 = \left(\sum_{i=1}^{k} \frac{T_i^2}{n_i}\right) - \frac{(\sum x)^2}{n}$
  - Between treatments mean square: $\text{MSTR} = \frac{\text{SSTR}}{k-1}$
  - Error sum of squares: $\text{SSE} = \sum_{i=1}^{k} (n_i - 1)s_i^2 = \sum_{i=1}^{k} \sum (x - \bar{x}_i)^2$
  - Error mean square: $\text{MSE} = \frac{\text{SSE}}{n-k}$
  - Total sum of squares: $\text{SST} = \text{SSTR} + \text{SSE} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$
- F ratio: $F = \frac{\text{MSTR}}{\text{MSE}} = \frac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)}$
  - $\nu_1 = k - 1$ is the degrees of freedom for the numerator
  - $\nu_2 = n - k$ is the degrees of freedom for denominator
  - F-distribution: $F(\nu_1, \nu_2)$
    * Mean of F-distribution: $\mu = \frac{\nu_2}{\nu_2 - 2}$
    * Let $F_\alpha(\nu_1, \nu_2)$ to be the critical value where α is the area at right hand tail, then we have $F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$
- If the means of the k populations are equal, then we would have $F < F_\alpha(\nu_1, \nu_2)$ with confidence $1 - \alpha$

11 Regression

11.1 Linear regression

- For every sample, we can obtain two properties X and Y, which we denote as (x_i, y_i) for sample i
- Sample size: n
- Whether X and Y are correlated? That is, if providing x_i, can we obtain projected y?

- Usually, assume X and Y satisfy the linear relation: $Y = a + bX$
- Then for known value $X = x_i$, compute the projected value $Y = \hat{y}_i$

- Sample mean:
  - Sample mean of X: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample mean of Y: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- Regression formula: For $Y = a + bX$,
  - $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $a = \bar{y} - b\bar{x}$
  - We can compute $\hat{y}_i = a + bx_i$

11.2 Correlation coefficient

- Sum of square error of regression: $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  which measures the difference between the experimental value and projected value of Y

- Spread of X: $\mathrm{SS}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}_i)^2$

- Spread of Y: $\mathrm{SS}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$

- Correlation of X and Y: $\mathrm{SS}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- Sum of square of regression variability: $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- Then we have:
  - $\mathrm{SS}_{yy} = \mathrm{SSE} + \mathrm{SSR}$
  - Proportion of error due to error spread= $\frac{\mathrm{SSE}}{\mathrm{SS}_{yy}}$

- Squared correlation: $R^2 = \frac{\mathrm{SSR}}{\mathrm{SS}_{yy}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SS}_{yy}} \leq 1$

- Correlation coefficient: $r = \mathrm{sign}(b)\sqrt{R^2}$
  - If $R^2 = 1$, then regression equation and experimental data fits exactly

- Pearson coefficient, a.k.a. product moment correlation coefficient: $r_{xy} = \frac{\mathrm{SS}_{xy}}{\sqrt{\mathrm{SS}_{xx} \cdot \mathrm{SS}_{yy}}}$
  - Same property as r