Lesson 1 Point Estimation

1.1 Definition

- Parameter Space

the range of possible values of the parameter, say, theta, is called the parameter space, denoted omega.

- Point Estimator

the statistic, u, a function of the random samples, is used to estimate the parameter theta and is called a point estimator of theta.

- Point Estimate

the function $u(x_1, x_2, x_3, \ldots, x_n)$ computed from a set of data is an observed point estimate of theta.

1.2 Maximum Likelihood Estimation

A good estimation of the unknown parameter theta would be the value of theta that maximizes the probability or likelihood of getting the data we observed.

- Definition

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \cdots, \theta_m$, with probability density or mass function $f(x_i; \theta_1, \theta_2, \cdots, \theta_m)$.

Suppose that $(\theta_1, \theta_2, \cdots, \theta_m)$ is restricted to a given parameter space omega. Then,

1. $$L(\theta_1, \theta_2, \cdots, \theta_m) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = f(x_1; \theta_1 \cdots \theta_m) \cdot f(x_2; \theta_1, \cdots, \theta_m) \cdot \cdots$$
$$= \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \cdots, \theta_m) \quad \text{where } (\theta_1, \theta_2, \cdots, \theta_m) \text{ in } \Omega.$$

It is called the likelihood function.

2. If $[u_1(x_1, x_2, \cdots, x_n), u_2(x_1, x_2, \cdots, x_n), u_m(x_1, x_2, \cdots, x_n)]$

is the m-tulle that maximizes the likelihood function, then $\hat{\theta}_i = u_i(x_1, x_2, \ldots, x_n)$

is the maximum likelihood estimator of θ_i.

3. The corresponding observed values of the statistics in (2), namely:

$$[u_1(x_1, x_2, \cdots, x_n),\ u_2(x_1, x_2, \cdots, x_n),\ u_m(x_1, x_2, \cdots, x_n)]$$

are called the maximum likelihood estimates of θ_i.

To solve the likelihood function and get the maximum value, one need to take the derivative and solve the equation after setting it to zero.

1.3 Unbiased Estimation

A good estimator is an unbiased estimator.

- Biased and Unbiased Estimator

if the mean of a statistic equals to the parameter, $E[\hat{\theta}_i] = \theta_i$, then we say the statistic is an unbiased estimator of the parameter theta.

Lesson 2 Confidence Intervals of One Mean

2.1 The Situation

Rather than using just a point estimation, we could find an interval or range of values that we can be really confident contains the actual value of unknown parameters.

An interval of such values, $L < \theta < U$, is called a confidence interval. Each interval has a confidence coefficient, $1-\alpha$, or a confidence level, $100(1-\alpha)\%$.

Typical confidence coefficients are 0.9, 0.95, and 0.99, with corresponding confidence levels 90%, 95%, and 99%. The interpretation is "we are 95% confident that the parameter falls between L and U".

2.2 A Z-Interval for a Mean

- Theorem

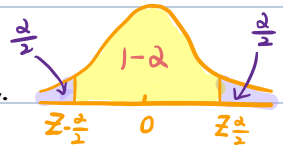
1. $X_1, X_2, \cdots, X_n$ is a random variable from a normal distribution with mean mu and variance, so that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

2. The population variance σ^2 is known. Then, a $(1-\alpha)100\%$ confidence interval for the mean μ is

$$\bar{X} \pm Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

The interval, because it depends on Z, is often referred to as the Z-interval for a mean.



2.3 Interpretation

It is incorrect to say that "the probability that the population mean falls between the lower value L and upper value U is $1-\alpha$". The reason is probability is about random variables. The population mean is a constant, not a random variable. It makes no sense to make a probability statement about a constant.

It is correct to think confidence interval in this way:

1. Suppose we take a large number of samples, say 1000,
2. Then, we calculate the 95% confidence interval for each sample,
3. Then, "95% confident" means that we would expect 95% of the 1000 intervals to be correct to contain the actual unknown value, mu.

2.4 An Interval's Length

Length = U-L, A narrower interval means a more specific range of the magnitude of the parameter.

When it is the case of the Z-interval, Length $= \left[\bar{x} + z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)\right] - \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)\right]$

As we can see, the interval length is affected by sample size n, variance, sigma, and confidence level.

So,

1. Increasing sample size n would shrink the length of the interval given other factors fixed
2. Decreasing the standard deviation sigma would shrink the length too, but we have no control over the population standard deviation
3. Decreasing the confidence level, the length of the interval decrease. But there is a trade off the length and how confident we are. 95% is used the most often.

Overall, increasing sample size is the most effective way to decrease the length of the interval.

2.5 A t-interval for a Mean

When we derive the formula for the Z-interval, we assume the standard deviation of the population is unknown. But in reality, we could never know the true variance or standard deviation of a population, which leads us to estimate the population standard deviation using the sample standard deviation.

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

Then, in deriving the confident interval using the true theta, we would have $\frac{\bar{x} - u}{S/\sqrt{n}}$.

Unlike $\frac{\bar{x}-u}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{\bar{x}-u}{s/\sqrt{n}}$ follows a t distribution.

- Theorem

If $x_1, x_2, \cdots, x_n$ are normally distributed with mean and variance, then $\frac{\bar{x}-u}{s/\sqrt{n}}$ follows a T distribution with n-1 degree of freedom. $\frac{\bar{x}-u}{s/\sqrt{n}} \sim t_{n-1}$

- Theorem

If $x_1, x_2, \cdots, x_n$ are normally distributed with mean and variance, then a $(1-\alpha)100\%$ confidence interval for the population mean u is

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

It is referred to as the t-interval for the mean.

*. $\bar{x}$ is a point estimate of mu

*. $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$ is an interval estimate of mu

*. $\frac{s}{\sqrt{n}}$ is the standard error of the mean

*. $t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$ is the margin of error

2.6 Non-Normal Data

It is helpful to note that as the sample size n increases, the T ratio $T = \frac{\bar{x}-u}{s/\sqrt{n}}$ approaches an approximate normal distribution regardless of the distribution of the original data.

Therefore, when n is large, usually greater than 30, $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$ and $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$ yields similar intervals.

In practice,

1. If the data are normally distribution, use $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$
2. If the sample size is large enough, greater than 30, $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$ and $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$ similar
3. If the data are not normally distributed and sample size is small, use $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$

Lesson 3 Confidence Intervals for Two Means

3.1 Two-Sample Pooled t-interval

- Theorem

If $X_1, X_2, X_3, \ldots, X_n \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, Y_3, \ldots, Y_n \sim N(\mu_Y, \sigma^2)$ are independent random samples,

then a $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$, the difference in the population means is:

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

where S_p^2, the "pooled sample variance":

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

is an unbiased estimator of the common variance σ^2.

Three assumptions are made in deriving the above confidence interval formula. They are

1. the random samples X_i and Y_i are independent
2. the random samples, X_i and Y_i, are from normal distributions
3. the populations of X_i and Y_i have the same variance

3.2 Welch's t-Interval

In the confidence interval deriving, we assume the variance of the two populations is equal. When the variances are not equal, we will use Welch's t-interval.

- Welch's t-interval

If the data are normally distributed or approximately normally distributed, and the population variances σ_X^2 and σ_Y^2 can't be assumed to be equal, then a $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, r} \cdot \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

where the r degree of freedom are approximated by

$$r = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

If necessary, r should take the integer portion of r.

3.3 Paired t-Interval

Paired t-interval is usually used to study the effect of a treatment. In general, when dealing with pairs of dependent measurements, we should use the sample mean difference, d, to estimate the difference in the values before and after the treatment. As long as the differences are normally distributed, we should use the $(1-\alpha)100\%$ t-interval for the mean,

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

Before	X_1	X_2	X_3	.....	X_n
After	Y_1	Y_2	Y_3	.....	Y_n
Diff	d_1	d_2	d_3	.....	d_n

$\bar{d} = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i$; $S_d = \sqrt{Var(d)}$

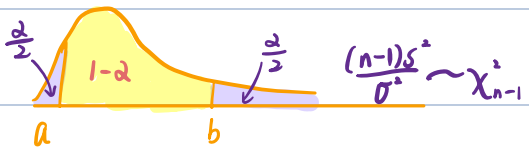
Lesson 4 Confidence Intervals for Variances

4.1 One Variance

- Theorem

If $X_1, X_2, X_3, \ldots, X_n$ are normally distributed and $a=\chi^2_{\frac{1-\alpha}{2},n-1}$, $b=\chi^2_{\frac{\alpha}{2},n-1}$, then $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is

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



Proof :

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

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

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

4.2 The F-Distribution

- F-Distribution

if U and V are independent chi-square random variables with r_1 and r_2 degrees of freedom, respectively, then $F = \frac{U/r_1}{V/r_2}$ follows an F-distribution with r_1 numerator degrees of freedom and r_2 denominator degrees of freedom. We write $F \sim F(r_1, r_2)$.

4.3 Two Variances

F-distribution helps us construct the confidence interval of the ratio of two population variances.

- Theorem

If $X_1, X_2, X_3, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, Y_3, \ldots, Y_n \sim N(\mu_y, \sigma_y^2)$ are independent random samples,

and $c = F_{\frac{1-\alpha}{2}}(m-1, n-1) = \frac{1}{F_{\frac{\alpha}{2}}(m-1, n-1)}$

$d = F_{\frac{\alpha}{2}}(m-1, n-1)$

Then a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_x^2}{\sigma_y^2}$ is : $\frac{1}{F_{\frac{\alpha}{2}}(m-1, n-1)} \cdot \frac{S_x^2}{S_y^2} \le \frac{\sigma_x^2}{\sigma_y^2} \le F_{\frac{\alpha}{2}}(m-1, n-1) \cdot \frac{S_x^2}{S_y^2}$

Lesson 5 Confidence Intervals for Proportions

5.1 One Proportion

When we study proportion questions, the sample proportion that employees want to WFH is 0.5, what is the confidence interval for the population proportion?

- Theorem

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

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

Note, we expect that $np >= 5$ and $n(1-p) >= 5$.

5.2 Two Proportions

Let's start to explore the confidence interval for the difference in two proportions. It is usually used to compare the proportions before and after some treatment.

- Theorem

For large random samples, an approximate $(1-\alpha)100\%$ confidence interval for p_1-p_2, the difference in two population proportions, is

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Lesson 6 Sample Size

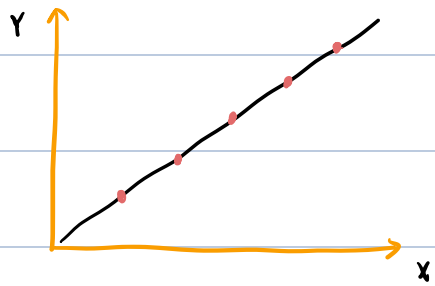
Usually, we want a sample size to bring us estimations under a certain error rate. We could determine the sample size based on our error tolerance using the general rule $n = \frac{(z^*_{\frac{\alpha}{2}}) \cdot s^2}{\varepsilon^2}$ for a mean inference and $n = \frac{(z^*_{\frac{\alpha}{2}}) \cdot p(1-p)}{\varepsilon^2}$ for a proportion inference.

Sample size is not usually a problem in most industries. I would like to skip this lesson and come back later when sample size is an issue.

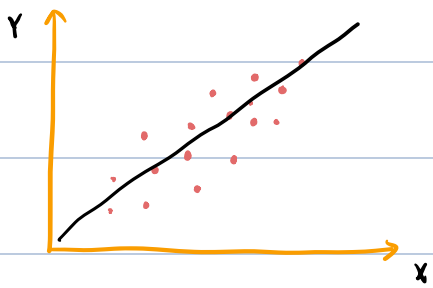
Lesson 7 Simple Linear Regression

7.1 Types of Relationship

- Deterministic relationships



- Statistical relationships



A regression problem is to study the statistical relationship between variables.

7.2 Least Squares: The Idea

To study the regression problem, we usually denote

Y_i: the i_{th} observed response

X_i: the i_{th} predictor values

$\hat{Y}_i$: the i_{th} predicted response or fitted value

e_i: $Y_i - \hat{Y}_i$, the prediction error or residual error

When we fit a line to best describe the statistical relationship like that in the plot of statistical relationship, we want the prediction error as small as possible. This idea is called least squares criterion". In short, the least squares criterion tells us in order to find the equation best fitting $\hat{y}_i = a + bx_i$

We need to choose the values a and b that minimize the sum of squared error $Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

7.3 Least Squares: The Theory

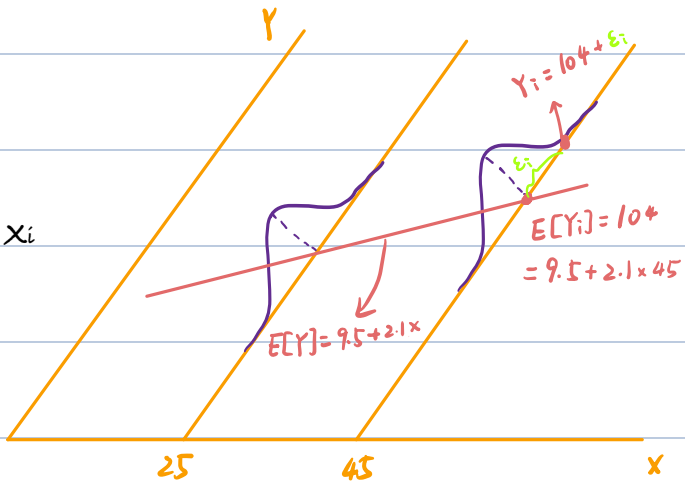
- Theorem

The least squares regression is: $\hat{y}_i = a + b(x_i - \bar{x})$ with least squares estimates

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

Considering the interpretation of a, using $(x_i - \bar{x})$ in the equation is more meaningful.

7.4 The Model



- Assumptions behind simple linear regression

1. the mean of the response, $E[Y_i]$, is a linear function of x_i
2. the errors, ε_i, are independent, so Y_i are independent
3. the errors, ε_i, are normally distributed, so Y_i do.
4. the errors, ε_i, have equal variance, so Y_i do, for all X values.

- Theorem

If the four assumptions hold true, then the least squares estimates are the same as these from maximum likelihood estimation.

- Theorem

The maximum likelihood estimator of the variance σ^2 is $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$

(It is an biased estimation of the variance)

- Mean Square Error

$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$ is an unbiased estimator of the variance.

7.5 Confidence Intervals for Regression Parameters

A traditional way to denote the parameters in simple linear regression is $a = \hat{\alpha}$, $b = \hat{\beta}$

- Theorem

Under the assumptions of simple linear regression model, $\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$

- Theorem

Under the assumptions of simple linear regression model, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$

- Theorem

Under the assumptions of simple linear regression model, $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$

- Theorem

Under the assumptions of simple linear regression model, a $(1-\alpha)100\%$ confidence interval for the slope parameter β is

$$\hat{\beta} \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{\frac{MSE}{\Sigma(X_i - \bar{X})^2}}$$

- Theorem

Under the assumptions of the simple linear regression model, a $(1-\alpha)100\%$ confidence interval for the intercept parameter α is

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

Lesson 8 More Regression

8.1 A Confidence Interval for the Mean of Y, E[Y]

- Theorem

A $(1-\alpha)100\%$ confidence interval for the mean μ_Y is $\hat{y} \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{MSE} \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}}$

8.2 A Prediction Interval for a New Y

- Theorem

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

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

Lesson 9 Tests about Proportions

Starting from here, it about hypothesis testing. Through hypothesis testing, we want to answer questions like: " is the value of the parameter theta such and such?"

Every time we perform a hypothesis test, this is the basic procedure:

1. Make an initial assumption about the population parameter
2. Collect evidence
3. Reject or fail to reject the initial assumption based on the evidence

There are two ways to making the decision:

1. Use critical value
2. P-value

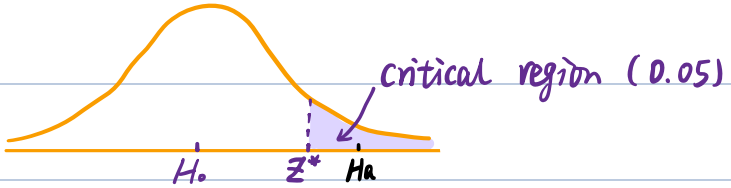
9.1 The Basic Idea

The tests about proportions Let us answers questions:

1. Whether a single population proportion p equals a particular value, p_0
2. Whether the difference in two population proportions p_1-p_2 equals a particular value, say, with the most common value being 0

- Basic Procedure for Testing Population Proportion ($\hat{P} \sim N(P_0, \frac{P_0(1-P_0)}{n})$

1. state the null hypothesis H_0 and the alternative hypothesis H_a.
2. Calculate the test statistics: $Z = \frac{\hat{P}-P_0}{\sqrt{P_0(1-P_0)/n}} \sim N(0,1)$
3. Determine the critical region or p-value
4. Make a decision, fail to reject or reject



*the test statistics fall in the critical region, we know null hypothesis has a small probability of happening. So, we reject it. O.W. we don't reject.

*The same as p-value. P-value is the probability that we observe a more extreme statistic if the null is true. If p-value is less than the significance level, alpha, we reject the null.

- Possible Errors

It is possible we make a wrong decision.

1. Type I Error

If we reject the null hypothesis when it is actually true, the error is called type I error. And the chance of making a type I error equals to the significance level of the test, usually, 0.05.

2. Type II Error

If we accept the null hypothesis when it is not true, the error is called Type II Error.

For example, we believe the probability of getting a 4 from a fair 4-sided die is 0.25. We use 0.25 as the population parameter in the calculation of Z statistics, which leads us to reject the null. But, if the actual probability is 0.27. Then, the Z statistics shows it is possible to accept the null. That means we make a type II error.

9.4 Comparing Two Proportions

Here, let's consider testing the equality of two proportions against the alternative that they are not equal. $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$

- Theorem

the test statistics for testing H_0 is $Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$; $\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$

Lesson 10 Tests about One Mean

10.1 Z-Test: When Population Variance is Known

- Hypothesis

$H_0: \mu = \mu_0$; $H_a: \mu < \mu_0$ (one-side) / $\mu \neq \mu_0$ (two-sided)

- Test Statistics

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Decision

Reject or not reject after comparing with z_α for one-sided test or $z_{\frac{\alpha}{2}}$ for two-sided test

10.2 T-Test: When Population Variance is Unknown

When population variance is unknown, the test statistics follows a T-distribution.

- Hypothesis

$H_0: \mu = \mu_0$ $H_a: \mu < \mu_0$ / $\mu > \mu_0$ / $\mu \neq \mu_0$

- Test Statistics

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

- Decision

Reject null or fail to reject null after comparing with $t_{\alpha,n-1}$ for a one-sided test or $t_{\frac{\alpha}{2},n-1}$ for a two-sided test.

10.3 Paired T-Test

There are occasions we are interested in comparing the means of two dependent population. E.g. we want to measure if a treatment takes effects or not by subtracting the measures before and after the treatment and test if the difference is significantly different from 0.

Then,

- Hypothesis

$H_0: \mu_d = \mu_x - \mu_y = \mu_0$; $H_a: \mu_d \neq \mu_0 \,/\, \mu_d > \,/\, \mu_d < \mu_0$; usually $\mu_0 = 0$

- Test Statistics

$$t = \frac{\bar{d} - \mu_0}{s_D / \sqrt{n}}$$

- Decision

Reject the null or do not reject the null after comparing with critical value or p-value to the significant level ; $t_{\frac{\alpha}{2}, n-1}$ (two-sided) ; $t_{\alpha, n-1}$ (one-sided) ;

Lesson 11 Test of the Equality of Two Means

A hypothesis test for the difference in two population means $\mu_1 - \mu_2$ has two situations:

1. when the unknown population variance are equal, it is a pooled two-sample t-test
2. When the unknown population variance are not equal, it is a Welch's t-test

11.1 When Population Variances Are Equal

If two independent samples from two normal distributions with equal variance $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then

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

where $S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$

- Ho

$$\mu_x - \mu_y = 0$$

- Ha

$$\mu_x \neq \mu_y \quad \text{or} \quad \mu_x > \mu_y \quad \text{or} \quad \mu_x < \mu_y$$

- Test Statistics

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

11.2 When Population Variance Are Not Equal

If the independent samples from two normal distributions with unequal variances $\sigma_x^2 \neq \sigma_y^2$, then

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim t_r \quad . \quad r = \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}$$

- Ho: $\mu_x - \mu_y = 0$
- Ha: $\mu_x - \mu_y \neq 0$ or $\mu_x > \mu_y$ or $\mu_x < \mu_y$

- Test Statistics: $T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$

Lesson 12 Tests for Variances

There are two topics for hypothesis tests for population variances:

1. A test for testing whether a single population variance σ^2 equals a particular value

2. A test for testing whether two population variances are equal

12.1 One Variance

If you have a random sample of size n from a normal distribution with mean μ and variance σ^2, then $$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Ho

$$\sigma^2 = \sigma_0^2$$

- Ha

$$\sigma^2 \neq \sigma_0^2 \quad \text{or} \quad \sigma^2 > \sigma_0^2 \quad \text{or} \quad \sigma^2 < \sigma_0^2$$

- Test Statistics

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

12.2 Two Variances

Suppose $X_1, X_2, X_3, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ_X and variance σ_X^2. And suppose $Y_1, Y_2, Y_3, \ldots, Y_m$ is from another random sample of size m and independent from X. It has mean μ_Y and variance σ_Y^2. Then,

$$F = \frac{\frac{(n-1)S_X^2}{\sigma_X^2} \Big/ n-1}{\frac{(m-1)S_Y^2}{\sigma_Y^2} \Big/ m-1} = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_X^2}{\sigma_Y^2} \sim F_{n-1,\, m-1}$$

- Ho: $\sigma_X^2 = \sigma_Y^2$
- Ha: $\sigma_X^2 \neq \sigma_Y^2$ or $\sigma_X^2 > \sigma_Y^2$ or $\sigma_X^2 < \sigma_Y^2$
- Test Statistics: $F = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_X^2}{\sigma_Y^2} = \frac{S_X^2}{S_Y^2}$

Lesson 13 One-Factor Analysis of Variance (ANOVA)

13.1 The Basic Idea

The analysis of variance is to compare the equality of the (unknown) means u_1, u_2, u_3, ⋯ , u_m of m normal distributions with an unknown but common variance σ^2 . This is the assumption of analysis of variance. It could be thought of as the extension of the pooled two-sample t-test.

- Ho

$$u_1 = u_2 = u_3 = u_4 = u_5 \cdots$$

- Ha

At least one of the mean differs from the others

- Test Statistics

We usually use statistics software to get the test statistics and p-value. We don't list the details of the mathematics here :)

- The Basic Idea Behind Analysis of Variance

Analysis of variance involves dividing the overall variability in observed data values so that we can draw conclusions of the populations from where the data come. The overall or total variability is divided into two components:

1. The variability "between" groups
2. The variability "within " groups

13.2 The ANOVA Table

One-Way Analysis of Variance Table Example

Source	DF	SS	MS	F	P
Factor	$m-1$	$SS(\text{Between})$	MSB	$\frac{MSB}{MSE}$	
Error	$n-m$	$SS(\text{Error})$	MSE		
Total	$n-1$	$SS(\text{Total})$			

- Source means "the source of the variance" in the data. It can be decomposed into factor, error, and total.
- DF means "the degrees of freedom in the source"
- SS means "the sum of squares due to the source"
- MS means "the mean sum of squared due to the source"
- F means "the F-Statistic"
- P means "the P-value"
- Factor means "the variability due to the factor of interest". It could be the variability due to different treatments, or the variability due to different methods. Sometimes, the row heading is labeled as Treatment. And sometimes, the row heading is labeled as Between to make it clear that the row concerns the variation between the groups.
- Error means "the variation within the groups" or "unexplained random error". Sometimes, the row heading is labeled as within to make it clear that the row concerns the variation within the group.
- Total means "the total variation in the data from the grand mean". That is, ignoring the factor of interest.
- $n-1$ means if there are n total data points collected, there are $n-1$ d.f.

- m-1 means if there are m groups being compared, then there are m-1 d.f. associated with the treatment/factor of interest
- n-m means if there are n total data points and m groups being compared, then there are n-m error d.f.
- SS(Between) is the sum of squares between group means and grand mean.
- SS(Error) is the sum of squares between the data points and the group means.
- SS(Total) is the sum of squares between the data points and the grand mean.
- MSB = SS(Between)/(m-1)
- MSE = SS(Error)/(n-m)
- F statistic is MSB/MSE. Because we want to compare the "average" variability between the groups to the average variability within the groups, we take the ratio of the MSB and MSE. Then, it follows a F distribution with $df_1 = m-1$ and $df_2 = n-m$.

Lesson 14 Two-Factor Analysis of Variance

One-way ANOVA tests the effect on one factor/treatment on the response variable. Two-way ANOVA is used to test the effect of two factors to understand whether either of the two factors or their interaction are associated with the response.

14.1 An Example

A study wants to learn how smoking history and types of stress tests effect the maximum oxygen uptake.

- Levels of smoking history: nonsmoker, moderate, heavy
- Types of stress test: bicycle, treadmill, step test
- Maximum oxygen uptake: continuous response in minutes
- Let alpha denote the effect of smoking, beta denote the effect of stress test, gamma denote the interaction between smoking and stress test.
- ANOVA Table

Source	DF	SS	MS	F	P	
Smoker	2	85	42	13	0.000	→ Reject $H_0: \alpha_1 = \alpha_2 = \alpha_3$ at $\alpha = 0.05$
Test	2	298	149	45	0.000	→ Reject $H_0: \beta_1 = \beta_2 = \beta_3$ at $\alpha = 0.05$
Smoker · Test	4	3	0.7	0.21	0.927	→ Fail to reject $H_0: \gamma_1 = \gamma_2 = \gamma_3$
Error	18	59	3			
Total	26	445				

- We can use the table to answer several hypothesis

1. Ho: $\gamma_1 = \gamma_2 = \gamma_3 = 0$
2. Ho: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ V.S. At least one is not equal to 0.

3. Ho : $\beta_1 = \beta_2 = \beta_3 = 0$

The assumption about the mean here is the group mean = $u_{ij} = u + \alpha_i + \beta_j + \gamma_{ij}$

Lesson 15 Tests Concerning Regression and Correlation

15.1 A Test for the Slope

In previous lessons, we have learned that $T = \frac{\hat{\beta} - \beta}{\sqrt{MSE / \Sigma(x_i - \bar{x})^2}} \sim t_{n-2}$

Then, a hypothesis test for the slope would be:

- Ho: $\beta = \beta_0$
- Ha: $\beta \neq \beta_0$ or $\beta > \beta_0$ or $\beta < \beta_0$
- Test statistic: $t = \frac{\hat{\beta} - \beta}{\sqrt{MSE / \Sigma(x_i - \bar{x})^2}}$

15.2 Three Tests for ρ

There are tests for the correlation coefficient, rho. It is not common in my work. So, I will skip these tests and will come back when it is needed.

Lesson 16 Chi-Square Goodness-of-Fit Tests

Starting from here, we are going to review nonparametric methods. In all previous tests, we assume our data follows a certain distribution. What if the assumptions don't hold? When this happens, alternatively, we could use nonparametric methods.

16.1 The General Approach

Goodness-of-Fit means how "good" the data "fit" the probability model. Suppose the student population in one school is 60% female and 40%male. If a sample of 100 students yields 53 females and 47 males. Can we conclude that the sample is random and representative of the population? That is, how "good" do the data "fit" the assumed probability model of 60% female and 40% male?

Testing whether there is a "good fit" between the observed data and the assumed probability model amounts to testing

$$H_0: P_F = 0.6 \quad \text{V.S.} \quad H_a: P_F \neq 0.6$$

Now, letting Y_1 denote the number of females selected, we know that Y_1 follows a binomial distribution with n trials and probability of success p_1. That is

$$Y_1 \sim b(n, p_1)$$

Then $E[Y_1] = np_1$, $Var(Y_1) = np_1(1-p_1)$

And, letting Y_2 denote the number of males selected, we know $Y_2 = n - Y_1$ follows a binomial distribution with n trials and probability of success p_2. That is,

$$Y_2 \sim b(n, p_2) = b(n, 1-p_1)$$

Then $E[Y_2] = np_2 = n(1-p_1)$, $Var(Y_2) = np_2(1-p_2) = n(1-p_1)(1-(1-p_1)) = np_1(1-p_1)$

Now, according to the central limit theorem, the normal is approximate to the binomial distribution for large samples $np \geq 5$ and $n(1-p) \geq 5$. Then, we have

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \overset{\text{approx.}}{\sim} N(0,1)$$

Therefore, squaring Z, we have

$$Z^2 = \left(\frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 = Q_1$$

$$\Rightarrow \quad Q_1 = Z^2 \cdot 1 = Z^2 \cdot ((1-p_1)+p_1)$$

$$= \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} \cdot ((1-p_1)+p_1)$$

$$= \frac{(Y_1 - np_1)^2 \cdot (1-p_1)}{np_1(1-p_1)} + \frac{(Y_1 - np_1)^2 \cdot p_1}{np_1(1-p_1)}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{((n-Y_2) - n(1-p_2))^2}{np_2}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

$$= \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i}$$

Observed → Expected

$$= \sum_{i=1}^{2} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \sim \text{Chi-Square, df} = 1$$

16.2 Extension to K Categories

The Chi-Square statistic that we derived on the previous page can be extended to accommodate any number of K categories.

- The Extension

Suppose an experiment can result in any of k mutually exclusive and exhaustive outcomes, A_1, A_2, A_3,...,A_n. If the experiment is repeated n independent times, an we let $p_i = P(A_i)$ and Y_i=the number of times the experiment results in A_i, $i=1,2,\ldots,k$, then we can summarize the number of observed outcomes and the number of expected outcomes for each of the k categories in a table as follows:

Categories	1	2	...	k-1	k
observed	Y_1	Y_2	...	Y_{k-1}	$n - Y_1 - Y_2 - \cdots - Y_{k-1}$
Expected	np_1	np_2	...	np_{k-1}	np_k

Karl Pearson showed that the Chi-Square statistic Q_{k-1} defined as

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$$

- Hypothesis

H0: $p_1 = 0.2$ $p_2 = 0.3$ $p_3 = 0.15$ $p_4 = 0.35$ ($k=4$ in this example)

Ha: p_i not specified in null

- Statistic

Q_{k-1}

- Conclusion

Reject the null if $Q_{k-1} > \chi^2_{k-1,\alpha}$

16.3 Unspecified Probabilities

In previous examples, we all assume we know the probability in the Binomial Distributions. What if the probabilities are not pre-specified? Chi-Square goodness-of-fit can still be used when parameters are unspecified.

If you are interested in testing whether a data set fits a probability model with d parameters left unspecified :

1. Estimate the d parameters using maximum likelihood method or other reasonable method.
2. Calculate the Chi-Square statistic Q_{k-1} using the obtained estimates.
3. Compare the Chi-Square statistic to a Chi-Square distribution with (k-1)-d degrees of freedom.

Lesson 17 Contingency Table

Using contingency table, we could

- Learn how to conduct a test for homogeneity
- Learn how to conduct a test for independence

17.1 Test for Homogeneity

The test for homogeneity is a method based on Chi-Square statistic, for testing whether two or more multinomial distributions are equal.

E.g. based on the data below, are females and males distributed equally among the various schools?

# Acceptance	Business	Engineer	Lib Arts	Science	Total → fixed!
Male	240 (20%)	480 (40%)	120 (10%)	360 (30%)	1200
Female	240 (30%)	80 (10%)	320 (40%)	160 (20%)	800
Total	480 (24%)	560 (28%)	440 (22%)	520 (26%)	2000

Then, what we want to test the hypotheses:

$$H_0: p_{MB} = p_{FB} \text{ and } p_{ME} = p_{FE} \text{ and } p_{ML} = p_{FL} \text{ and } p_{MS} = p_{FS}$$

$$H_a: p_{MB} \neq p_{FB} \text{ or } p_{ME} \neq p_{FE} \text{ or } p_{ML} \neq p_{FL} \text{ or } p_{MS} \neq p_{FS}$$

Before we calculate the test statistics, let us clarify some notation.

- Notation

The letter i will index the h row categories, the letter j will index the k column categories

# (ACC)	Bus ($j=1$)	Eng ($j=2$)	L Arts ($j=3$)	Sci ($j=4$)	(Fixed) Total
M ($i=1$)	y_{11} ($\hat{p}_{11}$)	y_{12} ($\hat{p}_{12}$)	y_{13} ($\hat{p}_{13}$)	y_{14} ($\hat{p}_{14}$)	$n_1 = \sum_{j=1}^{k} y_{1j}$
F ($i=2$)	y_{21} ($\hat{p}_{21}$)	y_{22} ($\hat{p}_{22}$)	y_{23} ($\hat{p}_{23}$)	y_{24} ($\hat{p}_{24}$)	$n_2 = \sum_{j=1}^{k} y_{2j}$
Total	$y_{11} + y_{21}$ ($\hat{p}_1$)	$y_{12} + y_{22}$ ($\hat{p}_2$)	$y_{13} + y_{23}$ ($\hat{p}_3$)	$y_{14} + y_{24}$ ($\hat{p}_4$)	$n_1 + n_2$

With

1. Y_{ij} denoting the number falling into the j^{th} category of the i^{th} sample
2. $P_{ij} = y_{ij}/n_i$ denoting the proportion in the i^{th} sample falling into the j^{th} category
3. $n_i = \sum_{j=1}^{k} y_{ij}$ denoting the total number in the i^{th} sample
4. $P_{ij} = (y_{1j} + y_{2j})/(n_1+n_2)$ denoting the (overall) proportion falling into the j^{th} category

Then, the Chi-square test statistic for testing the equality of two multinomial distributions is

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(y_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}$$

expected

observed

The statistic follows a chi-square distribution with k-1 degrees of freedom. Reject the null hypothesis of equal proportions if Q is large, that is $Q > \chi^2_{\alpha, k-1}$

How to understand this intuitively:

Suppose the college plans to accept n_1+n_2 new students in total. Then, $y_{11}+y_{21}$, $y_{12}+y_{22}$, $y_{13}+y_{23}$, and $y_{14}+y_{24}$ are the number of new students in each department. $\hat{P}_1 = (y_{11}+y_{21})/(n_1+n_2)$, $\hat{P}_2 = (y_{12}+y_{22})/(n_1+n_2)$, $\hat{P}_3 = (y_{13}+y_{23})/(n_1+n_2)$, and $\hat{P}_4 = (y_{14}+y_{24})/(n_1+n_2)$, are the proportion of new students of each department in total new students. Across all departments, suppose the total number of new male students they want to have are $y_{11}+y_{12}+y_{13}+y_{14}=n_1$, and the total number of new female students are $y_{21}+y_{22}+y_{23}+y_{24}=n_2$. The expected number of new male students and female students in Business department are $\hat{P}_1 \cdot n_1$ and $\hat{P}_1 \cdot n_2$. The expected number of new male students and female students in engineering department are $\hat{P}_2 \cdot n_1$ and $\hat{P}_2 \cdot n_2$. The expected number of new male students and female students in L Art department are $\hat{P}_3 \cdot n_1$ and $\hat{P}_3 \cdot n_2$. The expected number of new male students and female students in science department are $\hat{P}_4 \cdot n_1$ and $\hat{P}_4 \cdot n_2$. If the expected number of new students are similar to the observed number, then the difference between observed and expected would be small. The Q statistic would be small, too. Otherwise, the Q would be great and lead to the

rejection of the null hypothesis.

if we want to extend two categories to more than two, say h, the Q statistic would be

$$Q = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(y_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}$$

expected

observed

And it follows approximate chi-square distribution with (h-1)(k-1) of degrees of freedom.

- Sampling Scheme

1. taking h random and independent samples, with n_i fixed in advance

2. Observing into which of the k categories the each random sample fall

17.2 Test for Independence

Let start with an example again.

Suppose 395 people are randomly selected, and are "cross-classified" into one of eight cells, depending into which age category they fall and whether or not they smoke:

Smoke		Var B (Age)				
	observed	(18-24) B_1	(25-34) B_2	(35-49) B_3	(50-64) B_4	Total
Var A	(Yes) A_1	60	54	46	41	201
	(No) A_2	40	44	53	57	194
	Total	100	98	99	98	n=395

In this case, we are interested in conducting a test of independence to understand if smoking is associated with age. Then, the null hypothesis is

Ho: Variable A is independent of variable B, that is, $P(A_i)P(B_j) = P(A_i \cap B_j)$

Ha: Variable A is not independent of variable B

The sampling scheme involves:

1. Taking one random sample of size n, with n fixed in advance

2. Then "cross-classifying" each subject into one and only one of the mutually exclusive and exhaustive $A_i \cap B_j$ cells.

Note, in this case, only n is fixed, the row totals and column totals are random. That is the difference between test for homogeneity and test for independence.

- Notation

	Variable B				
Variable A	$B_1(j=1)$	$B_2(j=2)$	$B_3(j=3)$	$B_4(j=4)$	Total
$A_1(i=1)$	Y_{11} (P_{11})	Y_{12} (P_{12})	Y_{13} (P_{13})	Y_{14} (P_{14})	$P_{1\cdot}$
$A_2(i=2)$	Y_{21} (P_{21})	Y_{22} (P_{22})	Y_{23} (P_{23})	Y_{24} (P_{24})	$P_{2\cdot}$
Total	$P_{\cdot 1}$	$P_{\cdot 2}$	$P_{\cdot 3}$	$P_{\cdot 4}$	n

$\frac{Y_{11}+Y_{12}+Y_{13}+Y_{14}}{n}$

$\frac{Y_{14}+Y_{24}}{n}$

If variable A and variable B are independent, we would have $P_{ij} = (P_{i\cdot})(P_{\cdot j})$. $\hat{P}_{i\cdot} = \frac{\sum_{j=1}^{k} Y_{ij}}{n}$; $\hat{P}_{\cdot j} = \frac{\sum_{i=1}^{h} Y_{ij}}{n}$

Then, $\hat{P}_{ij} = \hat{P}_{i\cdot} \cdot \hat{P}_{\cdot j} = \frac{\sum_{j=1}^{k} Y_{ij}}{n} \cdot \frac{\sum_{i=1}^{h} Y_{ij}}{n} = \frac{Y_{i\cdot} \cdot Y_{\cdot j}}{n^2}$

The Q statistic is still formed by observed and expected value,

$$Q_{kh-1} = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(Y_{ij} - n\cdot P_{ij})^2}{n P_{ij}}$$

the expected value of $\hat{Y}_{ij} = n\cdot\hat{P}_{ij} = n\cdot\hat{P}_{i\cdot}\cdot\hat{P}_{\cdot j}$. Then,

$$Q_{kh-1} = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(Y_{ij} - n\cdot P_{ij})^2}{n P_{ij}} = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(Y_{ij} - n\cdot\hat{P}_{i\cdot}\cdot\hat{P}_{\cdot j})^2}{n\cdot\hat{P}_{i\cdot}\cdot\hat{P}_{\cdot j}}$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{\left(Y_{ij} - n\cdot\frac{Y_{i\cdot}\cdot Y_{\cdot j}}{n^2}\right)^2}{n\cdot\frac{Y_{i\cdot}\cdot Y_{\cdot j}}{n^2}}$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{\left(Y_{ij} - \frac{Y_{i\cdot}\cdot Y_{\cdot j}}{n}\right)^2}{\frac{Y_{i\cdot}\cdot Y_{\cdot j}}{n}} \quad \overset{\text{approx.}}{\sim} \quad \chi^2_{(h-1)(k-1)}$$

How to understand it intuitively:

It is based on the probability of independent events. We have row probability for variable A and column probability for variable B. Say, we have row probability of smoking=yes is 0.4, and column

probability of age group (18-24) 0.4. If smoking and age group are independent, then, the probability of being in this group and smoking is n*0.4*0.4=0.16n. But, if we observed the proportion in this cell is 0.5n , quite different from 0.16n. Then, our test maybe indicate the Q statistic much greater the criteria and reject the null, and tell us age group and smoking are not independent.At least, for age group(18-24), they have a higher probability of smoking.

Lesson 18 Order Statistics

Suppose, we would like to know the probability that the third largest value less than 72, we need to know how to order the data and understand the probability distribution of the ordered random samples. This is about order statistics.

18.1 The Basic

We have high chances having ties in our data. Here, we will assume that the n independent observations come from a continuous distribution, thereby making the probability zero that any two observations are equal.

- Definition

If $X_1, X_2, X_3, \dots, X_n$ are observations of a random sample of size n from a continuous distribution, we let the random variable: $Y_1, Y_2, Y_3, \dots, Y_n$ denote the order statistics of the sample, with :

1. Y_1 being the smallest of the $X_1, X_2, \dots, X_n$ observations
2. Y_2 being the second smallest of the $X_1, X_2, \dots, X_n$ observations
3. ...
4. Y_{n-1} being the next-to-largest of the $X_1, X_2, \dots, X_n$ observations
5. Y_n being the largest of the $X_1, X_2, \dots, X_n$ observations

Let's understand it with an example.

Let $Y_1<Y_2<Y_3<Y_4<Y_5<Y_6$ be the order statistics associated with $n=6$ independent observations each from the distribution with probability density function $f(x)=(1/2)x$ for $0<x<2$. What is the probability that the next-to-largest order statistics, that is, Y_5, less than 1? That is, what is $P(Y_5<1)$?

Answer

If we want $Y_5<1$, there are two situations. One is all 6 variables are less than one. The other is the first

5 variables are less than 1 and Y_6 greater than 1. It looks pretty similar, right? It is Binomial Distribution if we denote the success is a variable less than 1. And the number of success, denoted as Z, would be 6 for the first situation and 5 for the second situation. Now, we need to figure out the probability of the success. We have the density function of X, then, we could find the probability that $x<1$ from that.

$$P(X_i<1) = \int_0^1 f(x)\,dx = \int_0^1 \frac{1}{2}x\,dx = \frac{1}{2}\cdot\frac{1}{2}\cdot x^2\Big|_0^1$$
$$= \frac{1}{4}(1) - \frac{1}{4}(0)$$
$$= \frac{1}{4}$$

So, $Z \sim Bin(6, 0.25)$

Now, our question, $P(Y_5<1) = P(Z=5) + P(Z=6) = \binom{6}{5}\left(\frac{1}{4}\right)^5\cdot\binom{1}{1}\left(\frac{3}{4}\right)^1 + \binom{6}{6}\left(\frac{1}{4}\right)^6 = 0.0046$

A follow-up questions is what is the cumulative distribution function $G_5(y)$ of the order statistic Y_5?

$$G_5(y) = P(Y_5<y)$$
$$= P(Z=5) + P(Z=6) \quad \text{with } Z \sim Bin\left(6, \frac{y^2}{4}\right)$$
$$= \binom{6}{5}\left(\frac{y^2}{4}\right)^5\cdot\binom{1}{1}\left(1-\frac{y^2}{4}\right) + \binom{6}{6}\left(\frac{y^2}{4}\right)^6$$

Then, the next question is, what is the density function, $g_5(y)$ of the fifth order statistic Y_5?

$$g_5(y) = G_5'(y) \qquad f(x)=\frac{1}{2}x$$
$$= \frac{6!}{4!1!}\left(1-\frac{y^2}{4}\right)\left(\frac{y^2}{4}\right)^4\left(\frac{1}{2}y\right) \qquad F(x)=\frac{x^2}{4}$$
$$= \frac{6!}{4!1!}[1-F(y)]\cdot[F(y)]^4\cdot[f(y)]$$

18.2 The Probability Density Functions of Order Statistics

- Theorem

Let $Y_1 < Y_2 < \ldots < Y_{n-1} < Y_n$ be the order statistics of n independent observations from a continuous distribution with cumulative distribution function $F(x)$ and probability density function $f(x)$, where $0 < F(x) < 1$ over the support $a < x < b$. Then, the probability density function of the rth order statistics is

$$g_r(y) = \frac{n!}{(r-1)!(n-r)!}[F(y)]^{r-1}[1-F(y)]^{n-r} \cdot f(y)$$

over the support $a < y < b$.

Lesson 20 The Wilcoxon Tests

Most of the hypothesis testing procedures we have investigated so far depend on some assumptions about the distributions, like normality of the data. Distribution-free methods relays some of those distributional assumptions.

20.1 The Sign Test for a Median

A median, m, means 50% values are less than it and the other 50% are greater than it.

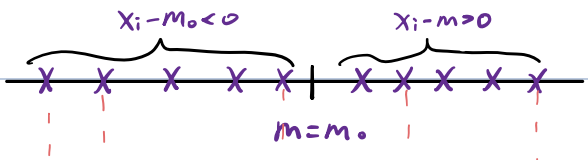
- Hypothesis

Ho: $m = m_0$

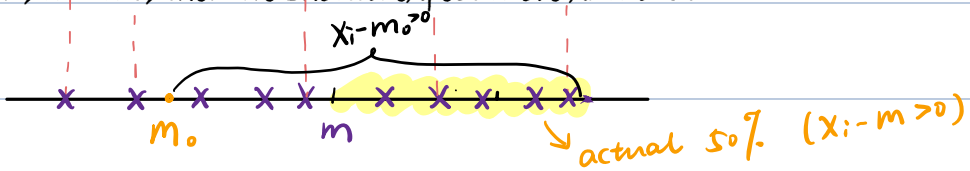
Ha: $m > m_0$ or $m < m_0$ or $m \neq m_0$

- Build the test

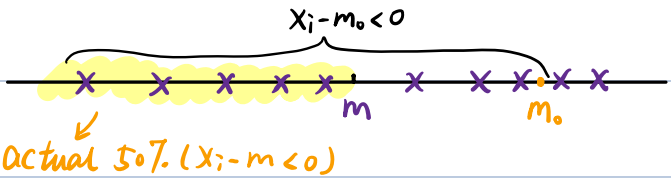
Let's start by considering the quantity $x_i - m_0$ for $i = 1,2,\ldots,n$ if the null hypothesis is true. Then, we should expect about half of the $x_i - m_0$ quantities obtained to be positive and half to be negative :



If instead, the actual median, $m > m_0$, then we should expect more $x_i - m_0 > 0$.



If, the actual median, $m < m_0$, then, we should have more negative values, that is $x_i - m_0 < 0$



The analysis of $x_i - m_0$ under the three situations $m = m_0$, $m > m_0$, and $m < m_0$ suggests then that a reasonable test for testing the value of a median m should depend on $x_i - m_0$. That is exactly what the

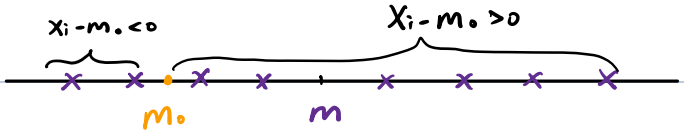
sign test for a median does. This is what we will do:

1. Calculate $x_i - m_0$ for $i = 1,2,\ldots,n$
2. Define N-, the number of negative signs obtained upon calculating $x_i - m_0$ for $i=1,2,\ldots,n$
3. Define N+, the number of negative signs obtained upon calculating $x_i - m_0$ for $i=1,2,\ldots,n$

Then, if the null hypothesis is true, that is, $m = m_0$, then N- and N+ both follow a binomial distribution with parameters n and p=0.5. That is $N- \sim Bin(n,0.5)$ and $N+ \sim Bin(n,0.5)$.

Then, we use the binomial distribution with n as the sample size to calculate the probability that we get the observed N- (observed negative signs) or N+, the observed positive signs. If the probability is pretty small and less than alpha, then we will reject the null hypothesis.

E.g. $H_0: m = m_0$ $H_a: m > m_0$



$P(n-) = P(N- < n-)$, reject H_0 if $P(n-) < \alpha$. (it means under this assumption, we have small probability to get the # of negative signs.)

↓ observed n- (# of $x_i - m_0 < 0$)

Note: If H_a is $m < m_0$, we would have more negative signs and few positive signs if H_a is true. Then we would calculate $P(n+) = P(N < n+)$

20.2 The Wilcoxon Signed Rank Test for a Median

It is like an extension of the sign test above and it introduces a new distribution, W distribution. I will not put details about this section as I am tired of typing all the functions 😴. I will review myself again in the future necessarily.

Lesson 21 Run Test and Test for Randomness

Run Test is used to test whether the distribution functions $F(x)$ and $G(y)$ of two continuous random variables X and Y are equal.

21.1 The Run Test

- What is a run?

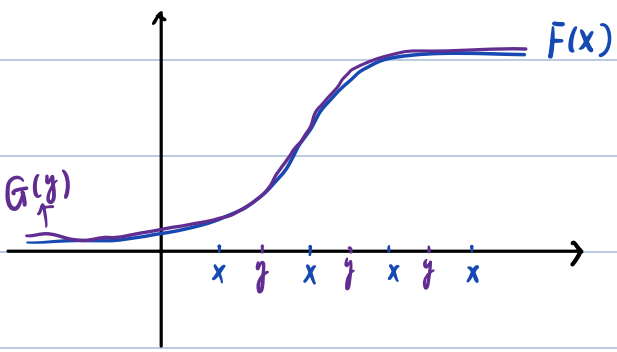
Let's suppose we have n_1 observations of random variable X_1 and n_2 observations of the random variable Y. Suppose we combine the two sets of independent observations into one large collection of n_1+n_2 observations, and then arrange the observations in increasing order of magnitude. If we label from which set each of the ordered observations originally come, we might observe something like this:

where x denote an observation of the random variable X and y denotes an observations of random variable Y. Then, each group of successive values of X and Y is what we call a run.

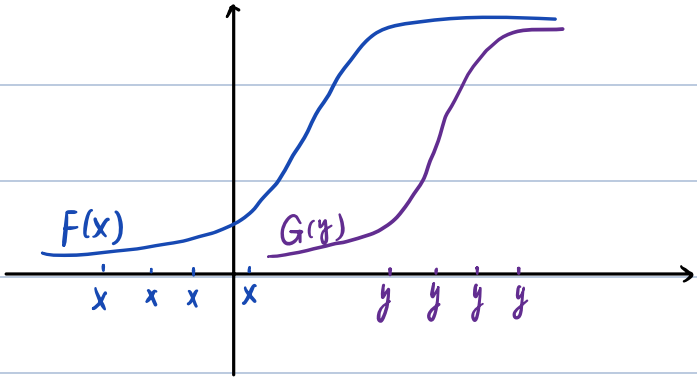
- Why runs?

Then, in what way, we might knowing the number of runs be helpful in testing the null hypothesis that $F(x)=G(y)$? Let's start with some examples.

1. if $F(x)=G(y)$, we could observe the plot below. Any value, Z, in the x-axis, would have $F(Z)=G(Z)$.



2. If $F(X) > G(y)$, we could observe a plot below. Any value, Z, in the x-axis, would have $F(Z) > G(X)$.



3. Another way of the inequality could look like the plot below.

In this case, the median of X and Y are nearly equal, but the variance of Y is much greater than the variance of X.

The above examples give a pretty clear indication that why we use runs to test the null hypothesis of the equality of two distribution functions. If the number of runs is smaller than expected, it seems we should reject the null because a small number of runs suggests that there are differences in either the location or the spread of the two distributions. To use the number of runs to conduct a testing, we need to know about the distribution of R, the number of runs.

- What is the p.m.f of R?

Let's use R to denote the number of runs in the combined ordered sample containing n_1 observations of X and n_2 observations of Y. R is discrete.

If the null hypothesis is true, the distribution function are equal. All the possible permutation of X's

and Y's in the combined sample are equally likely. Therefore, we can use the classical approach to assigning the probability that R equals a particular value r. That is, to find the distribution of R, we can find:

$$P(R=r) = \frac{\text{\# of ways of getting r runs}}{\text{total \# of ways of arranging x's and y's}}$$

for all of the possible values in the support of R.

The probability that R, the number of runs, takes on a particular value r, when r is even, is:

$$P(R=r=2k) = \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}$$

- The probability that R, the number of runs, takes on a particular value r, when r is odd, is:

$$P(R=r=2k+1) = \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_2-1}{k}\binom{n_1-1}{k-1}}{\binom{n_1+n_2}{n_1}}$$

- A Large-Sample Test

Conducting a single run test can be quite computationally expensive. There is an easier way to run the test providing n_1 and n_2 are large than 10.

If the samples are large, the distribution of the number of runs, R, can be approximated with a normal distribution. That is,

$$Z = \frac{R - \mu_R}{\sqrt{Var(R)}} \sim N(0,1)$$

$$\text{where} \quad \mu_R = \frac{2n_1n_2}{n_1+n_2} + 1, \quad Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

Because a small number of runs is evidence that the distribution functions are unequal, the critical region for testing the null hypothesis: $H_0: F(z) = G(z)$ is of the form $z \leq -z_\alpha$.

21.2 Test for Randomness

A common application of run test is a test for randomness of observations. Because an interest in randomness of observations is quite often seen in a quality-control setting, that is the application that will be the focus of our attention. For example, suppose a quality control supervisor at a paint manufacturing company suspects that the weights of paint cans on the production line are not varying in a random way, as she would expect. Then, she can run a test to determine if this is the case. Let's let $X_1, X_2, \ldots, X_N$ denote the observed weights, where the subscripts designate the order in which the outcomes were obtained. For the sake of concreteness, let's suppose they N, the total number of observations, is even,

. If we calculate the median m of the observed weights, then by definition, the median divides the observed sample in half. That is m half of the weights will be less than the median, and half of the weights will be greater than the median. Now, suppose we replace each observation by L if it falls below the median, and by U if it falls above the median.

1. If we observe something like this:

U U U U L U L L L L

Then, the production process is showing a trend. In this case, we would not observe as many runs r as we would expect if the process were truly random. In this case, we would reject the null hypothesis of randomness, in favor of the alternative hypothesis of a trend effect, if $r \leq c$.

2. If we instead observed something like this:

U L U L U L U L U L

Then, the production process is showing some kind of cyclic event. In this case, we would observe more runs r than we would expect if the process were truly random. In this case, we would reject the null

hypothesis of randomness, in favor of the alternative hypothesis of cyclic effect $r \geq c$.

3. If we aren't sure in which way a process would deviate from randomness, then we should allow for the possibility of either a trend effect or a cyclic effect. In this case, we should reject the null of randomness, in favor of the alternative hypothesis of either a trend effect or a cyclic effect, if $r \leq c$ or $r \geq c$

Lesson 22 Kolmogorov-Smirnoff Goodness-of-Fit Test

In this lesson, we will learn how to conduct a test to see how well a hypothesized distribution function $F(x)$ fits an empirical distribution function $F_n(x)$. In this lesson, we will

- learn a formal definition of an empirical distribution function
- Justify the Kolmogorov-Smirnov test statistic
- Try out the test on a few examples

22.1 The Test

- Empirical Distribution Function

Given an observed random sample $X_1, X_2, \ldots X_n$, an empirical distribution function $F_n(x)$ is the fraction of sample observations less than or equal to the value x. More specifically, if $y_1 < y_2 < \ldots < y_n$ are the order statistics of the observed random sample, with no two observations being equal, then the empirical distribution function is defined as:

$$F_n(x) = \begin{cases} 0, & \text{for } x < y_1 \\ \frac{k}{n}, & \text{for } y_k \le x \le y_{k+1}, \quad k = 1, 2, \cdots, n-1 \\ 1, & \text{for } x \ge y_n \end{cases}$$

That is, for the case in which no two observations are equal, the empirical distribution function is a step function that jumps $1/n$ in height at each observation x_k.

Example

A random sample of $n=8$ people yields the following (ordered) counts of the number of times they swam in the past month: 0, 1, 2, 2, 4, 6, 6, 7. Calculate the empirical distribution function $F_n(x)$.

Answer:

As the data are ordered, therefore the order statistics are

$y_1 = 0$, $y_2 = 1$, $y_3 = 2$, $y_4 = 2$, $y_5 = 4$, $y_6 = 6$, $y_7 = 6$ and $y_8 = 7$. Therefore, using the definition of the empirical distribution function, we have:

$$F_n(x) = 0 \quad \text{for} \quad x < 0$$

$$F_n(x) = \frac{1}{8} \quad \text{for} \quad 0 \le x < 1, \text{ as } y_1 = 0$$

$$F_n(x) = \frac{2}{8} \quad \text{for} \quad 1 \le x < 2, \text{ as } y_1 = 0, y_2 = 1$$

$$F_n(x) = \frac{4}{8} \quad \text{for} \quad 2 \le x < 3, \text{ as } y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 2$$

$$\dots$$

$$F_n(x) = 1 \quad \text{for} \quad x \ge 7$$

Now, let's jump right in and state and justify the Kolmogorov-Smirnoff statistic for testing whether an empirical distribution fits a hypothesized distribution well.

- Kolmogorov-Smirnov test statistic

$$D_n = \sup_x [|F(x) - F_0(x)|]$$

Is used for testing the null hypothesis that the cumulative distribution function $F(x)$ equals some hypothesized distribution function $F_0(x)$, that is, $H_0: F(x) = F_0(x)$, against all of the possible alternative hypotheses $H_a: F(x) \neq F_0(x)$. That is, D_n is the least upper bound of all pointwise difference $|F(x) - F_0(x)|$.

Therefore, at any point x, if there is a large difference between the empirical distribution $F_n(x)$ and the hypothesized distribution $F_0(x)$, it would suggest that the empirical distribution $F_n(x)$ does not equal the hypothesized distribution $F_0(x)$. Therefore, we reject the null hypothesis if D_n is too large. And we can use the lookup table to make the decision.

Lesson 23 Bayesian Methods

23.1 Subjective Probability

Here is an example that illustrates how a Bayesian might use available data to assign probabilities to particular events. The following amounts, in dollars, are bet on horses A, B, C and D to win a local race:

Horse	Amount
A	10000
B	30000
C	22000
D	38000
Total	100000

Determine the payout for winning with a $2 bet on each of the four horses, if the track wants to take 17%, or $17,000 from the amount collected.

Answer

The first thing we need to determine is the likelihood of winning. If $38,000 worth of bets came in for Horse D and only $10,000 came in for Horse A, that tells that more people think that Horse D is going to win than Horse A. It would behoove us to use the amounts bet to assign probabilities. That is,

the probability that Horse A will win is $P(A) = \frac{10000}{100000} = 0.1$

the probability that Horse B will win is $P(B) = \frac{30000}{100000} = 0.3$

the probability that Horse C will win is $P(C) = \frac{22000}{100000} = 0.22$

the probability that Horse D will win is $P(D) = \frac{38000}{100000} = 0.38$

Now, let's determine the odds of each winning. (Odds = P(failure)/P(success))

Odds against A = $\frac{1-0.1}{0.1} = 9$ to 1

Odds against B = $\frac{1-0.3}{0.3} = \frac{7}{3}$ to 1

Odds against C = $\frac{1-0.22}{0.22} = \frac{39}{11}$ to 1

Odds against D = $\frac{1-0.38}{0.38} = \frac{31}{19}$ to 1

The odds against A tells us that the track will payout $9 on a $1 bet;

The odds against B tells us that the track will payout $2.33 on a $1 bet;

The odds against C tells us that the track will payout $3.55 on a $1 bet;

The odds against D tells u that the track will payout $1.65 on a $1 bet.

Now, we can determine that the amount the track pays out on a $2 bet on each horse:

Payout for A = $2+9($2) =$20

Payout for B = $2+(7/3)($2) =$6.67

Payout for C = $2+(39/11)($2) =$9.09

Payout for D = $2+(31/19)($2) =$5.26

Now, we need to skim the 17% off for the race track to make sure that it makes some money off the bettors, multiplying each of the calculate payouts by 0.83 to get the final payout of each horse:

Absolute payout for A = $20(0.83) =$16.60

Absolute payout for B = $6.67(0.83) =$5.54

Absolute payout for C = $9.09(0.83) =$7.54

Absolute payout for D = $5.26(0.83) =$4.37

23.2 Bayesian Estimation

The key difference between frequentist statisticians and Bayesian statisticians is whether a statistian thinks of a parameter as some unknown constant or as random variable.

Example

A triffid control engineer believe that the cars passing through a particular intersection arrive at a mean rate , lambda, equal to either 3 or 5 for a given time interval. Prior to collecting any data, the engineer believes that it is much more likely that the rate lambda = 3 than lambda = 5. In fact, the engineer believes that the prior probabilities are $P(\lambda=3) = 0.7$, $P(\lambda=5) = 0.3$

One day, during a random selected time interval, the engineer observed $x=7$ cars pass through the intersection. In light of the engineer's observation, what is the probability that $\lambda=3$ and what is the probability that $\lambda=5$?

Answer

The problem is, given $x=7$, what is the probability that lambda=3, and given $x=7$, what is the probability that lambda=5.

Using Bayes' Theorem, we have

$$P(\lambda=3 \mid x=7) = \frac{P(\lambda=3, x=7)}{P(x=7)}$$

$$= \frac{P(\lambda=3) \cdot P(x=7 \mid \lambda=3)}{P(\lambda=3) \cdot P(x=7 \mid \lambda=3) + P(\lambda=5) \cdot P(x=7 \mid \lambda=5)}$$

We can use Poisson cumulative probability table to find the conditional probability in the denominator. From that, we have

$$P(x=7 \mid \lambda=3) = 0.22 \quad \& \quad P(x=7 \mid \lambda=5) = 0.105$$

Now, the desired probability is

$$P(\lambda=3 \mid x=7) = \frac{(0.7)(0.22)}{(0.7)(0.22) + (0.3)(0.105)} = 0.328$$

In this example, $P(\lambda = 3)$ is called the prior probability. That is because it is the probability that the parameter takes on a particular value prior to taking into account any new information. The new calculated probability, $P(\lambda = 3 | X = 7)$ is called posterior probability. That is because the probability that the parameter takes on a particular value posterior to taking into account the new information.

In this example, the parameter space of the parameter lambda only has 3 and 5. In other cases, where the parameter space for a parameter, say theta, takes on an infinite number of possible values, a Bayesian must specify a prior probability density function h(theta). Then, Bayesian estimation is used to find a posterior probability density function k(theta|y) if we know the probability density function g(y|theta).

So, if we know $h(\theta)$ and $g(y|\theta)$, we can treat the joint distribution of y and theta as

$$k(\theta, y) = g(y|\theta) \cdot h(\theta)$$

Then, the marginal distribution of y is

$$k(y) = \int_{-\infty}^{+\infty} k(\theta, y)\, d\theta = \int_{-\infty}^{+\infty} g(y|\theta) \cdot h(\theta)\, d\theta$$

And then, we can find the posterior distribution of theta given y using Bayes' Theorem, that is,

$$k(\theta|y) = \frac{k(\theta, y)}{k(y)} = \frac{g(y|\theta)h(\theta)}{k(y)}$$

Example

Suppose that Y follows a binomial distribution with parameter n and p = theta, so that the p.m.f. of Y given theta is

$$g(y|\theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y}$$

Suppose that the prior p.d.f of the parameter theta is the beta p.d.f., that is

$$h(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

for 0<theta<1. Find the posterior p.d.f of theta, given Y=y. And estimate theta using square error loss

Answer

1. find the joint distribution of theta and y

$$k(y,\theta) = g(y|\theta)\cdot h(\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{y}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}$$

2. find the marginal distribution of y

$$k(y) = \int_0^1 k(y,\theta)\,d\theta = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1\theta^{y+\alpha-1}\cdot(1-\theta)^{n-y+\beta-1}\,d\theta$$

$$= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\left(\frac{\Gamma(y+\alpha)\,\Gamma(n-y+\beta)}{\Gamma(y+\alpha+n-y+\beta)}\right)$$

3. find the posterior p.d.f. of theta <1st Question, Done!>

$$k(\theta|y) = \frac{k(y,\theta)}{k(y)} = \frac{g(y|\theta)\cdot h(\theta)}{k(y)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\,\Gamma(n+\beta-y)}\cdot\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

4. Estimate theta using square loss error function

The square loss error is $(\theta-w(y))^2$. We should use the conditional mean $w(y)=E(\text{theta}|y)$ as an estimate of the parameter theta. That is because we have shown in previous lessons that when $w(y)$ is the mean of the random variable would minimize the square loss.

$$E[\theta|y] = \frac{\alpha+y}{\alpha+y+n-y+\beta} = \frac{\alpha+y}{\alpha+n+\beta}$$