

STAT301: Cheat Sheet

- Algebra

- (i) $\frac{a+z \times b-a}{b} = z$ (ii) $a(b+c) = a \times b + a \times c.$
- (iii) $\frac{1}{\sqrt{ab}} = \frac{1}{\sqrt{a}} \times \frac{1}{\sqrt{b}}.$
- (iv) $\frac{a}{\sqrt{a}} = \sqrt{a}.$
- (v) $a < b$ means a is *less* than b . $a > b$ means a is *bigger* than b . $a \leq b$ means that a is *less than or the same* as b .
 - The point that cuts the interval $[a, b]$ in half is $\frac{(a+b)}{2}.$

- Probability The chance of a certain event happening.

Example: Out of 44 calves, 12 weighed less than 90 pounds. The probability of randomly picking a calf from this group which weighs less than 90 pounds is 12/44.

- Types of studies

- Observational Record data on individuals without attempting an intervention.
- Experimental Deliberately impose a treatment on individuals. Usually this is done in a randomized fashion, where some are given a treatment and others a placebo.

- Confounding When a observed factor and unobserved factor are mixed-up, making it impossible to decide what is influencing the response.

- Types of variable Numerical discrete, numerical continuous and categorical.

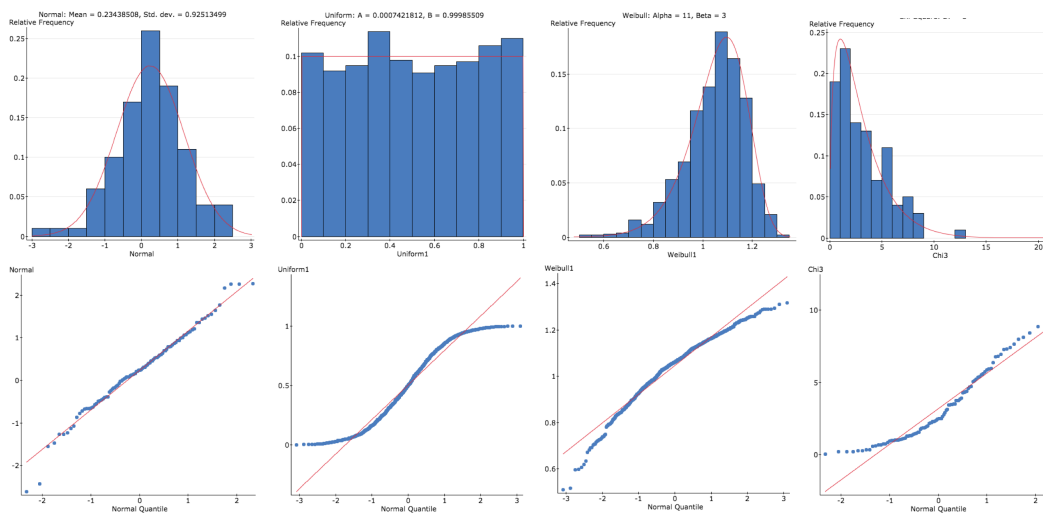


Figure 1: Shapes of distribution: Symmetric, Uniform (thick/heavy tailed), Left Skewed, Right Skewed and corresponding QQplots

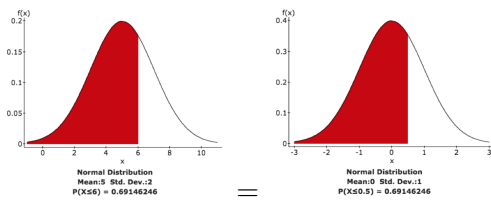
Data Analysis

- Checking for normality

- Use QQplot, see Figure 1.

- A cruder method is to use the 68-95-99.7% rule (check to see whether the data is within one, two and three sample sd of the sample mean).
- Measures of center given data X_1, \dots, X_n
 - (i) (Sample) mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.
Example: Average of 1, 1, 1, 3, 4, 5, 6 is $\bar{x} = 3$
 - (ii) Median is the point which cuts the data in half.
Example: Median of 1, 1, 1, 3, 4, 5, 6 is 3. Median of 1, 1, 3, 4, 5, 6 is 3.5.
- Measure of spread of given data X_1, \dots, X_n .
 - (i) (Sample) standard deviation $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$.
Example: The standard deviation of 1, 1, 1, 3, 4, 5, 6 is $s = 2.08$.
 - (iii) Quartiles and Interquartile range The first quartile cuts the first half of the data in half and the second quartile cuts the top part of the data in half. IQR = 3rd quartile - 1st quartile.
Example: The first and third quartile of 1, 1, 1, 3, 4, 5, 6 is 1 and 5 respectively.
- Linear transformation Suppose the data X_1, \dots, X_n has sample mean \bar{X} and sample standard deviation s_X . We make a linear transformation of the data set using the transformation $Y_i = a + bX_i$. The sample mean and standard deviation of the new data set is $\bar{Y} = a + b\bar{X}$ and $s_Y = |b|s_X$ respectively.
Example: 0.5, 1.5, 2, 3.2, 3.8 has mean 2.2 and standard deviation 1.3. We transform it using $Y = 12X$. The new data set is 6.0, 18.0, 24.0, 38.4, 45.6 which has mean 12×2.2 and standard deviation 1.3×12 .
- Z-score calculations Suppose X is a random variable with mean μ and standard deviation σ , the z-transform is $Z = \frac{X - \mu}{\sigma}$. The mean and standard deviation of the z-transform is zero and one.
The z-transform tells us how many standard deviations an observation, X , is from the mean μ .

Normal distribution

- Normal calculations. Note the normal distribution is (i) symmetric about the mean (ii) total area is one (iii) the y-axis is positive.
 - Question: Suppose the random variable X is known to come from a normal distribution with mean 5 and standard deviation 2 $N(5, 2)$. What is the chance X will be less than 6?
 - Answer: Make z-transform $z = \frac{6-5}{2} = 0.5$ then look up 0.5 (from outside into the z-tables) to give $P(X \leq 6) = P(Z \leq 0.5) = 0.69$:
- 
- Question: Suppose that X is known to come from a normal distribution with mean 5 and standard deviation 2 $N(5, 2)$. If an observation X is in the 85th percentile what is X ?
 - Answer: Look up 0.85 (from inside to outside) the table, which corresponds to 1.04, so $X = 5 + 1.04 \times 2 = 7.08$.
 - Rule of Thumb: If data is normally distributed, the roughly speaking 68% of the data lies within one standard deviation of the mean, 95% of the data lies within two standard deviations of the μ mean and 99.8% of the data lies within 3 standard deviations of the mean.

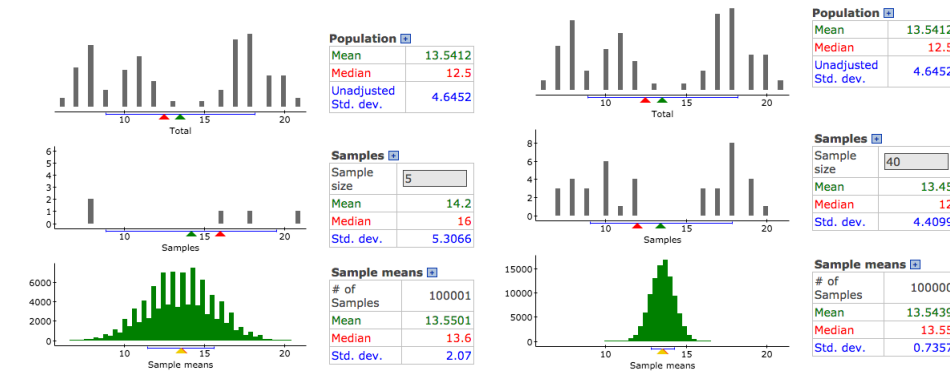


Figure 2: The distribution of averages

The sample mean

- The sample mean Suppose a random sample X_1, \dots, X_n is drawn from a population, where the mean is μ and the standard deviation is σ . The average, usually called the sample mean, $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \sum_{i=1}^n X_i$ is an estimator of the sample mean.
- Mean and standard error of the sample mean
 - The mean of the sample mean (the average of the average) is μ .
 - The standard error (variability) of the sample mean is σ/\sqrt{n} . The standard error informs us how variable the estimator is. The smaller the standard error, the less variable it will be.
 - Example: A population has mean $\mu = 67$ and standard deviation $\sigma = 3.8$. A sample of 5 is drawn and the average is taken. The average will change from sample to sample, but it is estimating the population mean. The mean of the average, \bar{X} , (the average of the average) is again $\mu = 67$ (it is estimating this value, so it unbiased) and the standard error of the average, \bar{X} , is $s.e = \sigma/\sqrt{n} = \frac{3.8}{\sqrt{5}}$.
- The distribution of the sample mean
 - Normal data If the distribution of the population is normal (examples include heights of one gender) then the distribution of the sample mean (no matter how big or small the sample size) will be normal.
 - Example 1: Female heights are normally distributed with mean 64 inches and standard deviation 2.5 inches ($N(64, 2.5)$). A sample of size three is taken the average is $N(64, \frac{2.5}{\sqrt{3}})$.
 - Example 2: Female heights are normally distributed with mean 64 inches and standard deviation 2.5 inches ($N(64, 2.5)$). A sample of size 50 is taken the average is $N(64, \frac{2.5}{\sqrt{50}})$.
 - Non-normal data If the distribution of the population is not normal (examples include the number of M&Ms in a bag) then the distribution of the sample mean will be close to normal if the sample size is sufficiently large. How large is large depends on how close to normal the original distribution.
 - Example 1: The mean number of M&Ms in a bag is $\mu = 13.54$ with standard deviation $\sigma = 4.64$. The average in 5 bags of M&Ms will have mean $\mu = 13.54$ and standard error $se=4.26/\sqrt{5}$, but it will NOT be normally distributed because the original data is not normal.
 - Example 2: The mean number of M&Ms in a bag is $\mu = 13.54$ with standard deviation $\sigma = 4.64$. The average in 5 bags of M&Ms will be close to normal with $N(13.54, 4.26/\sqrt{40})$.
- If the sample mean is close to normal we can use all the usual normal calculations (using the mean and standard error) to calculate probabilities.

Inference for the sample mean

- Confidence Intervals A confidence interval is an interval where we believe with $C\%$ confidence the population mean lies. Typically $C = 95\%, 99\%, 90\%$. To construct a confidence interval using the sample mean \bar{X} (which is evaluated from the data) we need to be sure that the sample mean is normally distributed (either by normality of the data or the sample size being large enough for the CLT to kick in). We consider the two cases, which depends on whether the population standard deviation is known or not.

- (i) **Known population standard deviation σ** If for some reason the population standard deviation is known but the population mean is unknown the the 95% CI for the mean is $\left[\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}\right]$ (we look up 2.5% in the z-tables to get 1.96).
- (ii) **unknown population standard deviation σ** If the population standard deviation is unknown then we need to *estimate* it from the data. If the sample size is n , we replace the normal distribution with the t -distribution with $(n - 1)$ degrees of freedom. The 95% CI for the mean is $\left[\bar{X} \pm t_{n-1}(2.5\%) \times \frac{s}{\sqrt{n}}\right]$ (remember we need to look up 2.5% each side). As the sample size grows the difference between the normal and the t -distribution becomes less.

Example The sample size is 30, the sample mean is 0.5 and sample standard deviation $s = 4$, the 95% CI is $[0.5 \pm 2.04 \times 4/\sqrt{30}]$.

- Margin of Error This is half the length of the confidence interval.

Example The margin of error of 95% CI [3, 8] is $\text{MoE} = (8 - 3)/2$.

- Formula for Margin of Error If the population standard deviation is known, then the MoE for a 95% CI is $1.96 \times \frac{\sigma}{\sqrt{n}}$. We can use this to find the minimum sample size to obtain a given margin of error: $n = (1.96 \times \sigma/\text{MoE})^2$. Notes:

- The larger the standard deviation σ the larger the sample size we will need.
- If the standard deviation is unknown then bounds are given say, its somewhere between σ_1 to σ_2 . Use the largest standard deviation to get the smallest margin of error.
- To decrease the margin of error from m to m/P you need to increase the sample size by a factor P^2 .

- Testing the mean Depending on what the alternative of interest is, there are three different possible test set-ups. To reduce algebra we will assume the mean under investigation is 5.

- $H_0 : \mu = 5$ against $H_A : \mu \neq 5$.
- $H_0 : \mu \leq 5$ against $H_A : \mu > 5$.
- $H_0 : \mu \geq 5$ against $H_A : \mu < 5$.

Which hypothesis you use depends on the alternative that you want to ‘prove’.

Example: The mean height of females 30 years ago was known to be 63 inches. It is believed that female heights have increased over the past 30 years, what is the hypothesis of interest? Answer $H_0 : \mu \leq 63$ against $H_A : \mu > 63$.

- Let us suppose that X_1, \dots, X_n (these are numbers) is a random sample of size n , drawn from a population with mean μ (this is what we are investigating) and standard deviation σ . We will assume that the sample size is large enough such that the sample mean is normally distributed with mean μ and standard error σ/\sqrt{n} .

If the population standard deviation is unknown and is instead estimated from the data, then in all the calculation use a t -distribution with $n - 1$ -degrees of freedom rather than the standard normal distribution.

- Calculating the p-value The p-value is always calculated under the null. This means determining the chance of the observations if the null were true (how viable is the null)?

Example 1: We test the hypothesis $H_0 : \mu = 5$ against $H_A : \mu \neq 5$. We collect a random sample of size 30, the sample mean based on this sample is $\bar{X} = 6$ and the sample standard deviation is $s = 3$.

The t-transform is $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{6-5}{3/\sqrt{30}} = 1.825$. To calculate the p-value:

- 1 Calculate the smallest area under the plot, in this case it is the area to the RIGHT of 1.825. Using t-tables with 29df, we see that it is between 2.5-5%.
2. The p-value for the **two**-sided test, is **two** times this area, which is between 5-10%.

Example 2: We test the hypothesis $H_0 : \mu \leq 5$ against $H_A : \mu > 5$. We collect a random sample of size 30, the sample mean based on this sample is $\bar{X} = 6$ and the sample standard deviation is $s = 3$.

The t-transform is $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{6-5}{3/\sqrt{30}} = 1.825$. To calculate the p-value:

- 1 Check to see the direction of the alternative. Since $H_A : \mu > 5$, the alternative is pointing RIGHT.
2. The p-value for this **one**-sided test, is the area to the RIGHT of $t = 1.825$. From tables this area, is between 5-10%.
3. For the one-sided test the p-value is this area, which is between 5-10%.

Example 3: We test the hypothesis $H_0 : \mu \geq 5$ against $H_A : \mu < 5$. We collect a random sample of size 30, the sample mean based on this sample is $\bar{X} = 6$ and the sample standard deviation is $s = 3$.

The t-transform is $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{6-5}{3/\sqrt{30}} = 1.825$. To calculate the p-value:

- 1 Check to see the direction of the alternative. Since $H_A : \mu < 5$, the alternative is pointing LEFT.
2. The p-value for this **one**-sided test, is the area to the LEFT of $t = 1.825$. From tables this area, is between 90-95%.
3. For the one-sided test the p-value is this area, which is between 90-95%.

- The decision process The decision process is made at the α (typically 5%) significance level. We reject the null and say there is evidence to suggest the alternative is true (or equivalently there is evidence to reject the null), if the p-value is **less** than $\alpha\%$.

α is often called the type I error (or significance level). The larger α the more likely we are to falsely reject the null when the null is true.

Example In a tomato packing plant, the mean weight of tomato boxes is tested at the $\alpha\%$ significance level, every hour. If the machine is working correctly for every 100 tests, on average we will falsely reject (determine the machine faulty) α times.

- Confidence intervals and p-values The $(100 - \alpha)\%$ (eg. 95%) confidence interval and a test done at the $\alpha\%$ ($\alpha\%$) significance level are connected in the sense that bounds for p-values can be deduced from the confidence interval (this is because the length of the confidence interval and the non-rejection region are the same).

Example The 95% CI for the mean is $[0.5, 4]$. The sample mean is $\bar{X} = (4 + 0.5)/2 = 2.25$. Using this we can deduce the following:

1. Two-sided tests

- (a) We test the hypothesis $H_0 : \mu = 0$ against $H_A : \mu \neq 0$. Since 0 is not in the interval $[0.5, 4]$ the p-value for this two sided test is less than 5%. This means the smallest area is less than 2.5% and is the area to the RIGHT of $\bar{X} = 2.25$ (centered about zero).
- (b) We test the hypothesis $H_0 : \mu = 1$ against $H_A : \mu \neq 1$. 1 is inside the interval $[0.5, 4]$, thus it is plausible that the mean is 1. The p-value for this two-sided test is greater than 5%. This means that p-value for the smallest area is greater than 2.5% and is the area to the RIGHT of $\bar{X} = 2.25$ (centered about one).

2. One-sided test pointing RIGHT

We test the hypothesis $H_0 : \mu \leq 0$ against $H_A : \mu > 0$. The p-value is the area to the RIGHT of $\bar{X} = 2.25$ (centered about zero), which we have shown in (a) is LESS than 2.5%.

We test the hypothesis $H_0 : \mu \leq 1$ against $H_A : \mu > 1$. The p-value is the area to the RIGHT of $\bar{X} = 2.25$ (centered about one), which we have shown in (b) is GREATER than 2.5%.

3. One-sided test pointing LEFT

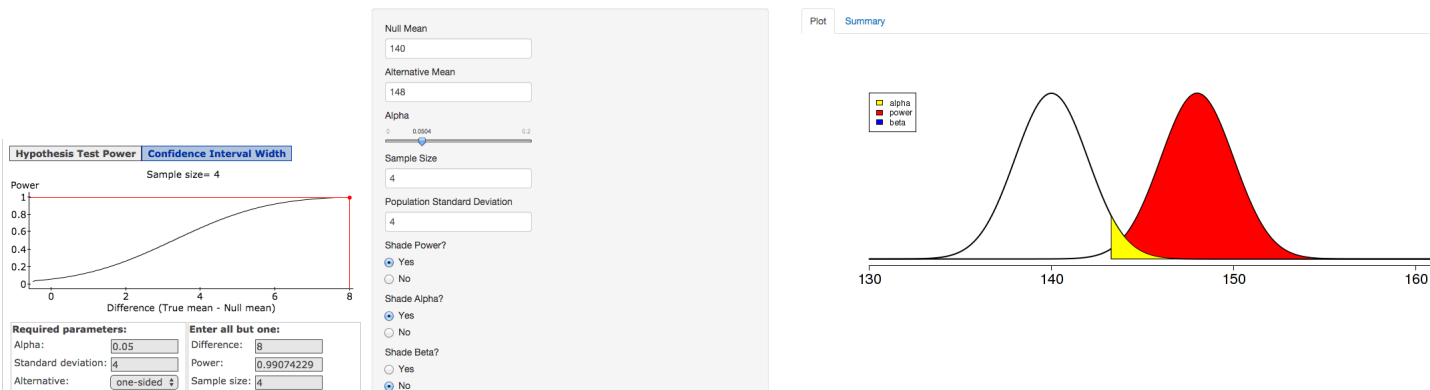
We test the hypothesis $H_0 : \mu \geq 0$ against $H_A : \mu < 0$. The p-value is the area to the LEFT of $\bar{X} = 2.25$ (centered about zero), which we have shown in (a) is GREATER than 97.5%.

We test the hypothesis $H_0 : \mu \geq 1$ against $H_A : \mu < 1$. The p-value is the area to the LEFT of $\bar{X} = 2.25$ (centered about one), which we have shown in (b) is less than 97.5%. But since $\bar{X} = 2.25$ is on the right of 1 it has to be greater than 50%. Thus the p-value is between 50 to 97.5%.

- Type I, Type II errors and Power Whenever we do a test we always base the result on a pre-set significance level (typically 5% usually denoted as α). These are features about the statistical procedure and not the data itself.

Example, suppose we test for gestational diabetes and use the hypothesis $H_0 : \mu \leq 140$ against $H_A : \mu > 140$.

- Type I error is the same as the significance level $=\alpha\%$ and is pre-set by us. What it means: If we set $\alpha = 5\%$, then the proportion of healthy women with $\mu = 140$, that we would falsely diagnose as having gestational diabetes is 5%.
- Type II error Given an alternative of interest the type II error is the chance of rejecting the null when actually that alternative is true. What it means: If someone is said to have severe diabetes when their mean is 148 then the type II error is the probability our test does not detect her. It can be calculated using: We see that when the sample size is 4 and the test is done at



the 5% level the chance of this happening is $1 - 0.9907 = 0.0093\%$.

- Power The power is the chance of rejecting null when the alternative is true. Using the above plot we see the power for the alternative $\mu = 148$ is 99.07%.

- Increasing the type I error increases power. Increasing sample size increases power.

Comparing populations

One of the most common statistical methods is when samples are collected from two different populations and compared (for example does a drug work verses a placebo). Hence there are two different data sets. How we analyze the data depends on how the data was collected.

Matched paired t-procedure:

- This a method for analyzing two data-set where there appears to be a clear matching/pairing in the data. If there is, a difference between the data is take and a one sample method (as described above) is applied to these differences (note because there is matching the size of the two samples have to be the same).

Example: Nutritionist want to investigate where red wine increases polyphenol levels in healthy males. To do a random sample of of 15 healthy males were taken (before the start of the experiment none drank red wine on a regular basis). Their polyphenol level was take before they started the experiment, their polyphenol level was measure. For two weeks they drank one glass of red wine each night. At the end of two weeks their level was taken again. The results of the test $H_0 : \mu_A - \mu_B \leq 0$ against $H_A : \mu_A - \mu_B > 0$ is give below. We see from the output that there is evidence to suggest

Before	After	After - Before
42.7	43.4	0.7
43.6	47.1	3.5
42.9	46.9	4
43.3	48.2	4.9
36.8	42.3	5.5
38.6	45.6	7
37.3	44.7	7.4
37.9	46	8.1
41.9	50.3	8.4
43	46.2	3.2
40.2	41	0.8
32.4	36.7	4.3
31.8	31.6	-0.2
36.7	36.1	-0.6
45	52.5	7.5

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
After - Before	4.3	0.79162281	14	5.4318799	<0.0001

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
After - Before	4.3	0.79162281	14	5.4318799	<0.0001

that drinking red wine increases polyphenol levels. Observe the t-value $= \frac{4.3}{0.79} = 5.4$. The 95% CI for the difference in the means $[4.3 \pm 2.14 \times 0.79]$ (where 2.14 was found using the t-distribution with 14df with 2.5%).

The independent (two) sample t-test:

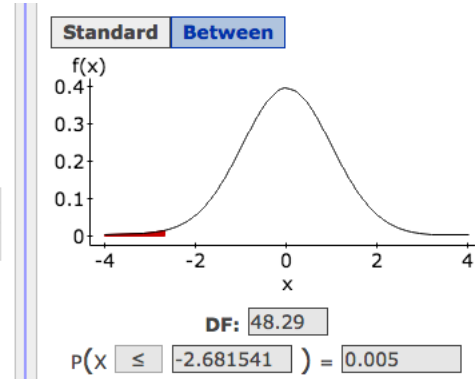
- If the two samples are take from two populations and there is no dependence between the two samples (no relationship), then an independent (two) sample procedure must be done. Because there is no matching between the two samples the sample sizes can be different.

Example: It is believed that male students at A&M are taller than female students at A&M. We do this by comparing their population means. The hypothesis of interest is $H_0 : \mu_M - \mu_F \leq 0$ against $H_A : \mu_M - \mu_F > 0$. The results of the test is given below. Note $t = \frac{0.466-0}{0.064} = 7.27$. The same output can be used to deduce the 95% confidence interval (if the 2.5% value is given).

Hypothesis test results:

μ_1 : Mean of males
 μ_2 : Mean of females
 $\mu_1 - \mu_2$: Difference between two means
 $H_0 : \mu_1 - \mu_2 = 0$
 $H_A : \mu_1 - \mu_2 > 0$
 (without pooled variances)

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	0.46665666	0.064165466	48.292554	7.2727073	<0.0001



Proportions

In this section we consider responses which are 'binary', eg. Yes or No. Guess or not Guess, Vote for Candidate B not vote for candidate B. We take a random sample of 'individuals' and use this to estimate the proportion of the population who say 'Yes', Gussed or vote for candidate A.

- One sample proportions Suppose the proportion of successes (this can be anything) 'population' is p and a random sample of size n is drawn from the sample. The number of success in a sample of size n follows what is known as a Binomial distribution $Bin(p, n)$. The Binomial is NOT the same as a normal distribution, however as the sample size n grows the 'closer' it gets to normality (this is the central limit theorem), the size of the sample depends on how close to p is to half (the further p is from $1/2$ the more skewed the distribution and the larger a sample size is required for the central limit to set in). We can use these results to make tests on proportions and construct confidence intervals (do not mix p and \hat{p} with p-value!).

Example Out of a random sample of 140 people 90 said they supported Gay Marriage. Based on this data, is there any evidence that the majority of the population support Gay Marriage?

The hypothesis of interest in this case is $H_0 : p \leq 0.5$ against $H_A : p > 0.5$. If the null were true the distribution of numbers out of 140 is given in the binomial distribution below. We see from both

Hypothesis test results:

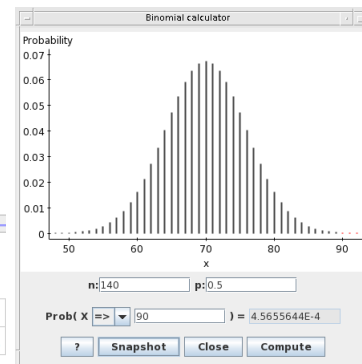
p : Proportion of successes
 $H_0 : p = 0.5$
 $H_A : p > 0.5$

Proportion	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
p	90	140	0.64285714	0.042257713	3.380617	0.0004

95% confidence interval results:

p : Proportion of successes
 Method: Standard-Wald

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	90	140	0.64285714	0.040496194	0.56348606	0.72222822



the Statcrunch crunch test and the binomial calculator that the p-value is extremely small (less than 0.4%), thus there is strong evidence against null, suggesting that majority do support Gay-marriage.

The slight discrepancy is because the Statcrunch output uses a normal approximation, which is calculated by making the z-transform

$$z = \frac{0.6428 - 0.5}{0.0422} = 3.38$$

and using the z-tables to calculate the area to the RIGHT of 3.38.

The 95% confidence interval for the location of the mean is $[0.6428 \pm 1.96 \times 0.0405] = [0.56, 0.72]$, which tells us that with 95% confidence the proportion that do support Gay marriage lies in this interval.

- Margin of Errors and Sample sizes The standard error for proportions $\sqrt{\frac{p(1-p)}{n}}$. Thus the margin of error (for a 95% confidence interval) and sample size for a given margin of error is

$$MoE = 1.96\sqrt{\frac{p(1-p)}{n}} \quad \text{and sample size} \quad n = \frac{1.96^2 p(1-p)}{MoE^2}.$$

Of course, p is unknown. Thus we need to choose the p which maximises the variability. If p is completely unknown then choose $p = 1/2$ (since $p(1-p)$ is maximised when $p = 1/2$). If some bounds on p are given, say it is known that p is in a given range, then choose the p in the range which is closest to $1/2$. Eg. if it is known that p lies in the interval $[0, 0.3]$, then use $p = 0.3$ in the sample size calculation. A plot of standard errors is given in Figure 3.

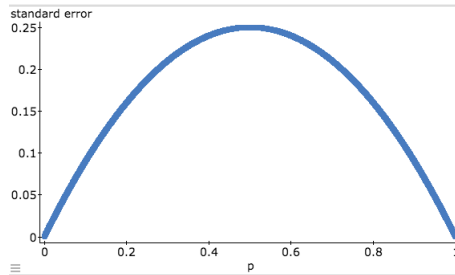


Figure 3: Standard error for a given sample size