1 Introduction to Probability and Statistics

1.1 Introduction

If you've heard this story before, don't stop me, because I'd like to hear it again. *Groucho Marx*

Many questions have probabilistic solutions. For example, are two objects/motifs similar because of some relationship, or is it coincidence? How do you classify similar items? How surprising is the occurence of an unusual event? How do we describe probabilistic outcomes using models? What do we do when an exact model is not available? Which hypothesis/model is more relevant/accurate/believable?

Before we talk models, we need to know about discrete and continuous random variables, and the testing of hypotheses. *Random* experiments must satisfy the following:

Random experiments

1. All possible discrete outcomes must be known in advance.
2. The outcome of a particular trial is not known in advance, and
3. The experiment can be repeated under identical conditions.

A sample space S is associated with a random experiment = set of all outcomes.

Example 1.1. S for a coin toss is $\{H, T\}$, and for a die roll $= \{1, 2, 3, 4, 5, 6\}$. An event A is a subset of S, and we are usually interested in whether A happens or not. e.g. $A =$ odd value on roll of a die; $A = H$ on tossing a coin.

The random variable X is some function mapping outcomes in space S to the real line. $X : S \to \mathbb{R}$. e.g. with two coin tosses, $S = \{(H, T), (T, H), (H, H), (T, T)\}$, $X =$ number of heads turning up, $X(H, T) = 1$, $X(H, H) = 2$ etc.

Random variable

Random variable: A real valued function assigning probabilities to different events in a sample space. This function could have discrete or continuous values. For example,

1. Sum of two rolls of a die ranges from $2 - 12$
2. The sequence $HHTH$ of coin tosses is not a random variable. It must have a numerical mapping.

Such variables are i.i.d.: independent and identically distributed.

A function of a random variable defines another random variable. Measures such as 'mean' and 'variance' can be associated with a random variable. A random variable can be conditioned on an event, or on another random variable. The notion of independence is important: the random variable may be independent from an event or from another random variable.

"Random variable" $\neq$ "No control over outcome".

Statistics refers to drawing inferences about specific phenomena (usually a random phenomenon) on the basis of a limited sample size. We are interested in descriptive statistics: we need measures of location and measures of spread.

Data scales are classified as categorical, ordinal or cardinal. *Categorical scale*: A label per individual (win/lose etc.). If the label = name, we refer to the data as nominal data. *Ordinal data* can be ranked, but no arithmetic transformation is meaningful. e.g. vision (= 20-20, 20-30), grades etc. *Cardinal data* is available in interval or ratio scales. *Ratio scale*: we know zero point (Eg. ratio of absolute temperatures is meaningful in K and not °C. The same for mass flow ratios.) *Interval scale (arbitrary zero)*: °F, °C.

We mostly use cardinal data.

Cardinal data may be sorted and ranked: info may be lost in the process, but analysis may be easier. Consider the following transformation: If 4 individuals A, B, C & D have heights (in inches) 72, 60, 68 and 76 respectively, they may be transformed to ranks: 2, 4, 3 & 1. The height difference between B & A $\neq$ difference between B and C, and therefore transformations may lead to incorrect inferences. Ordinal data are protected by monotonic transforms: If $x > y$, then $\log(x) > \log(y)$; however $|x - y| \neq |\log(x) - \log(y)|$ in general.

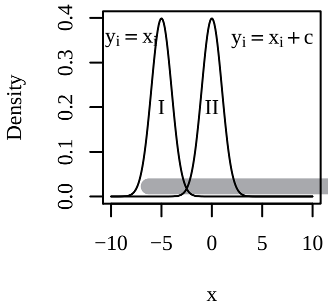


Fig. 1.1: Translating variables on a linear scale.

1.2 Discrete random variables

Measures of location

Mean

The sample mean is $\bar{x}$ and the population mean is μ.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1}$$

where N = size of total population and is usually unknown. The arithmetic mean can be translated or rescaled.

Translation: If c is added to every x_i, then if $y_i = x_i + c$, then $\bar{y} = \bar{x} + c$ (Fig. 1.1).

Scaling: If $y_i = cx_i$ then $\bar{y} = c\bar{x}$.

Median

For n observations, the sample median is

- $(n+1)/2$th largest observation for odd n,
- average of $n/2^{\text{th}}$ and $(n/2)+1^{\text{th}}$ largest observations for even n.

Mode

Most frequently occurring value. Can be problematic if each x_i is unique.

Geometric mean

$\bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$ and on a log scale, antilog$[(\sum_{i=1}^{n} \log x_i)/n]$

Harmonic mean

$1/\bar{x}_H = (1/n)\sum_{i=1}^{n}(1/x_i)$. For positive data, $\bar{x}_H < \bar{x}_G < \bar{x}$.

Display of data

Pareto diagram

Dot diagram

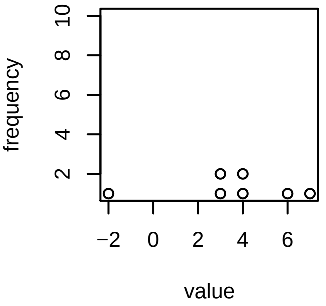


Fig. 1.2: The dot diagram.

Given $n = \{3, 6, -2, 4, 7, 4, 3\}$ the dot diagram is shown in Figure 1.2. This simple plot can be useful if there exists lots of data. From a visual inspection of the plot, -2 could be considered an outlier. Another advantage of these plots is that they can be superimposed.

Frequency Distributions

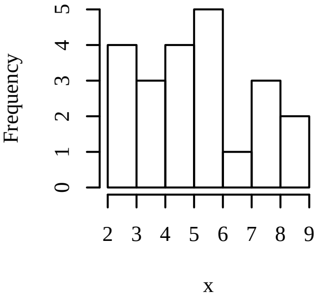


Fig. 1.3: A frequency distribution.

A frequency distribution is shown as a histogram in Figure 1.3. Frequencies for each class/category are tallied (sometimes using the # notation). Endpoints can be conveniently represented using bracket notation: in $[1, 10)$, 1 is included and 10 is not. In $[10, 19)$ 10 is included when tallying. Grouping together data loses information (in the form of specific individual values). Cumulative distributions can also be generated from the frequency curves. If the height of a rectangle in a histogram is equal to the relative frequency/width ratio, then total area = 1 ⇒ we have a density histogram.

Stem and leaf displays

1	2	7	5				12	17	15		
2	9	1	5	4	8		29	21	25	24	28
3	4	9	2	4		i.e.	34	39	32	34	
4	4	8	2				44	48	42		
5	3						53				

Measures of spread

Range

The range is the difference between largest and smallest observations in a sample set. It is easy to compute and is sensitive to outliers.

Quantiles/percentiles

The p^{th} percentile is that threshold such that p % of observations are at or below this value. Therefore it is the

- $(k+1)^{th}$ largest sample point if $np/100 \neq$ integer. [k = largest integer less than $np/100$].
- Average of $(np/100)^{th}$ and $(np/100+1)^{th}$ values if $np/100 =$ integer. For 10^{th} percentile $p/100 = 0.1$

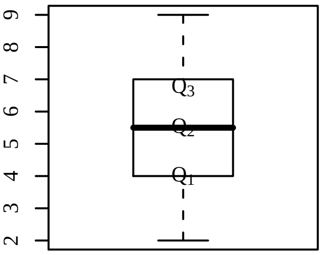


Fig. 1.4: A box plot.

The median = 50^{th} percentile. The layout of quantities which calculation of percentiles requires gives information about spread/shape of distribution. However data must be sorted.

Let Q_1= 25^{th} percentile, Q_2= 50^{th} percentile = median, and Q_3 = 75 th percentile. Then for the p^{th} percentile, (1) order the n observations (from small to large), (2) determine $np/100$, (3) if $np/100 \neq$ integer, round up, (4) if $np/100 =$ integer $= k$, average the k^{th} and $(k+1)^{th}$ values.

The range = (min value - max value) was strongly influenced by outliers. The interquartile range = $Q_3 - Q_1$ = middle half of data and is less affected by outliers. The boxplot (Fig. 1.4) is a graphical depiction of these ranges. A box is drawn from Q_1 to Q_3, a median is drawn as a vertical line in the box, and outer lines are drawn either up to the outermost points, or to 1.5×(box width) = 1.5×(interquartile range).

Deviation

The sum of deviations from the mean is $d = \sum_{i=1}^{n}(x_i - \overline{x})/n = 0$ and is not useful. The mean deviation is $\sum_{i=1}^{n}|x_i - \overline{x}|/n$. This metric does measure spread, but does not accurately reflect a bell shaped distribution.

Variance and Standard deviation

A variance may be defined as

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{1.2}$$

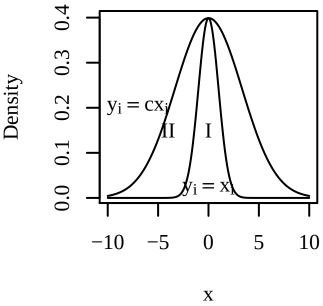


Fig. 1.5: Effect of scaling on variance.

where N denotes the size of a large population. The sample variance S^2 uses $(n-1)$ instead of n in the denominator. (There are only $(n-1)$ independent deviations if $\overline{x}$ is given, with their sum being constrained to be 0). This definition of S^2 with $(n-1)$ is used to better represent the population variance σ^2. A large S^2 implies large variability. S^2 is always > 0. The definition of S^2 above requires that $\bar{x}$ be computed first and then S^2; two passes of the data are required. A more convenient form (given below) permits simultaneous computation of $\sum x_i^2$ and $\sum x_i$, and thus requires one pass.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - (\sum x_i)^2/n}{n-1} \tag{1.3}$$

For translated data, $y_i = x_i + c$ ($i = 1$ to n) $\Rightarrow \overline{y} = \overline{x} + c$ and $S_y^2 = S_x^2$. For scaled samples, $y_i = cx_i$ for $i = 1, ..., n$, $c > 0 \Rightarrow S_y^2 = c^2 S_x^2$ (Fig. 1.5).

Coefficient of variation

$$CV = 100\% \frac{S}{\overline{x}} \tag{1.4}$$

This metric is dimensionless, hence one can discuss S relative to the mean's magnitude. This is useful when comparing different sets of samples with different means, with larger means usually having higher variability.

Indices of diversity

Such indices are used for nominal scaled data where mean/median do not make sense e.g. to describe diversity among groups. Uncertainty is a measure of diversity. For $k =$ categories and $p_i = f_i/n =$ proportion of observations in a category

More on uncertainty later. For now, notice the connection to entropy!

$$H = -\sum_{i=1}^{k} p_i \log p_i = \frac{n \log n - \sum_{i=1}^{k} f_i \log f}{n} \tag{1.5}$$

$H_{max} = \log k$. Shannon's index $J = H/H_{max}$.

Probability mass function (probability distribution)

The assignment of a probability to the value r that a random variable X takes is represented as $Pr_X(X = r)$ Eg. Two coin tosses $X =$ number of heads.

$$\Rightarrow Pr_X(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/4 & \text{if } x = 0 \text{ or } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Obviously $0 \leq Pr(X = r) \leq 1$ and $\sum_{r \in S} P(X = r) = 1$.

Notation: we will use $P(X = r)$ instead of $Pr_X(X = r)$

A probability distribution is a model based on infinite sampling, identifying the fraction of outcomes/observations in a sample that should be allocated to a specific value. Consider a series of 10 coin tosses where the outcome = 4 heads. A probability distribution can tell us how often 4 heads crop up when the 10 coin toss experiment is repeated 1000 times.

To establish the probability that a drug works on $\{0, 1, 2, 3, 4\}$ out of 4 patients treated, theoretical calculations (from some distribution) are compared with some frequency counts generated from 100 samples.

r	$P(X = r)$	Frequency distribution (100 samples)
0	0.008	0.000 [0/100]
1	0.076	0.090 [9/[100]
2	0.265	0.240 [24/100]
3	0.411	0.480 [48/100]
4	0.240	0.190 [19/100]

Compare the two distributions. Is clinical practice ↔ expectations from experience? Given data and some understanding of the process, you can apply one out of many known PMF's to generate the middle column above.

Expectation of a discrete random variable

The expectation of X is denoted

$$\mathcal{E}\left[X\right] \equiv \mu = \sum_{i=1}^{R} x_i P(x_i) \tag{1.6}$$

r	$P(X = r)$
0	0.008
1	0.076
2	0.265
3	0.411
4	0.240
	1.000

where x_i = value i of the random variable. e.g. probability that a drug works on $\{0, 1, 2, 3, 4\}$ out of four patients. $\mathcal{E}\left[X\right] = 0(0.008) + 1(0.076) + 2(0.265) + 3(0.411) + 4(0.240) = 2.80$. Therefore 2.8 out of 4 patients would be cured, on average.

For the case of two coin tosses, if X = number of heads, then $\mathcal{E}\left[X\right] = 0(1/4) + 1(1/2) + 2(1/4) = 1$. If the probability of a head on a coin toss was $3/4$

$$P(X = k) = \begin{cases} (1/4)^2 = 1/16 & k = 0 \quad \text{Binomial random variable} \\ 2 \times 1/4 \times 3/4 & k = 1 \qquad n = 2,\, p = 3/4 \\ (3/4)^2 & k = 2 \end{cases}$$

$\mathcal{E}\left[X\right] = 0(1/4)^2 + 1(2 \times 1/4 \times 3/4)^2 + 2(3/4)^2 = 24/16 = 1.5$

Expectations can be compared.

Example 1.2. Consider a game where 2 questions are to be answered. You have to decide which question to be answer first. Question 1 may be correctly answered with probability 0.8 in which case you win Rs. 100. Question 2 may be correctly answered with probability 0.5 in which case you win Rs. 200. If the first question is incorrectly answered, the game ends. What should you do to maximize your prize money expectation?

Soln:

Notice a tradeoff: Asking the more valuable question first runs the risk of not getting to answer the other question. If total prize money = X = random var, then $\mathcal{E}\left[X\right]$ assuming $Q1$ is asked first is $PMF[X] : p_X(X = 0) = 0.2$, $p_X(X = 100) = p_X(X = 300) = 0.8 \times 0.5$. Therefore $\mathcal{E}\left[X\right] = 0 \times 0.2 + 100 \times 0.8 \times 0.5 + 300 \times 0.8 \times 0.5 =$ Rs. 160.

$\mathcal{E}\left[X\right]$ assuming $Q2$ is asked first is $PMF[X] : p_X(X = 0) = 0.5$, $p_X(X = 100) = 0.5 \times 0.2$, $p_X(X = 300) = 0.5 \times 0.8$ hence $\mathcal{E}\left[X\right] = 0 \times 0.5 + 200 \times 0.5 \times 0.2 + 300 \times 0.5 \times 0.8 =$Rs. 140. Hence ask question 1 first.

Expectation comparisons can be misleading!

Example 1.3. The weather is good with a probability of 0.6. You walk 2 km to class at a speed $V = 5$ kph or you drive 2 km to class at a speed $V = 30$ kph. If the weather is good you walk. what is the mean of the time T to get to class? What is the mean of velocity V?

Soln:

$$\text{PMF of } T = P_T(t) = \begin{cases} 0.6 & \text{if} \quad t = 2/5 \text{ hrs} \\ 0.4 & \text{if} \quad t = 2/30 \text{ hrs} \end{cases}$$

Therefore $\mathcal{E}\left[T\right] = 0.6 \times 2/5 + 0.4 \times 2/30 = 4/15$ hrs. $\mathcal{E}\left[v\right] = 0.6 \times 5 + 0.4 \times 30 = 15$ kph. Note that $\mathcal{E}\left[T\right] = 4/15 \neq 2$ km$/\mathcal{E}\left[v\right] = 2/15$.

$T = 2/V$. However $\mathcal{E}\left[T\right] = \mathcal{E}\left[2/V\right] \neq 2/\mathcal{E}\left[V\right]$.

Variance of a discrete random variable

The variance of a discrete random variable is defined as

$$\operatorname{Var}\left[X\right] = \sigma_X^2 = \sum_{i=1}^{R} (x_i - \mu)^2 P(X = x_i) \tag{1.7}$$

The standard deviation $SD(X) = \sqrt{\operatorname{Var}\left[X\right]}$ = and has the same units as the mean (and X). Using the $\mathcal{E}\left[\right]$ notation,

$$\operatorname{Var}\left[X\right] = \sigma_X^2 = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])^2\right] = \mathcal{E}\left[X^2\right] - (\mathcal{E}\left[X\right])^2 = \mathcal{E}\left[X^2\right] - \mu^2 \tag{1.8}$$

Moments

μ'_k denotes the k^{th} moment about the origin

$$\mu' = \sum_i x_i^k P_X(x_i) \tag{1.9}$$

Hence the mean $= \mu = \mathcal{E}\left[X\right] = \sum_i x_i P(x_i) = 1^{st}$ moment about origin. μ_k is the k^{th} moment about mean.

$$\mu_k = \sum_i (x_i - \mu)^k P_X(x_i) \tag{1.10}$$

Variance $= 2^{nd}$ moment about mean $= \mu_2$. Then $\mu_2 = \mu_2^{'} - \mu^2$

$\mu_3/\sigma^3 =$ skewness (degree of asymmetry).

$$\frac{\mu_3}{\sigma^3} = \frac{\mathcal{E}\left[(X - \mathcal{E}\left[X\right])^3\right]}{\sigma^3} \tag{1.11}$$

Skew to right ($+ve$ skew) or left ($-ve$ skew). The mean is on same side of mode as longer tail. An alternate measure of skew is (mean - mode)/σ, but to avoid using mode, use skewness = 3×(mean - median)/σ.

Kurtosis (peakedness) is

$$\frac{\mu_4}{\sigma^4} = \frac{\mathcal{E}\left[(X - \mathcal{E}\left[X\right])^4\right]}{\sigma^4} \tag{1.12}$$

A moment generating function of X is defined as

If t is replaced with $-s$ you have the Laplace transform, and if replaced with it, the Fourier transform. It's a small world...

$$M(t) = M_X(t) = \mathcal{E}\left[e^{tX}\right] = \mathcal{E}\left[1 + tX + \frac{t^2x^2}{2!} + ...\right] \tag{1.13}$$

Then $\mu'_n = \mathcal{E}\left[X^n\right]$ is the n^{th} moment of X and is equal to $M^{(n)}(0)$ (from a comparison with a Taylor's expansion of $M(t)$), the n^{th} derivative of M, evaluated at $t = 0$.

$$M(t) = 1 + \mu'_1 t + \frac{\mu'_2 t^2}{2!} + ... + \frac{\mu'_n t^n}{n!} + ...$$

If $y = \sum_{i=1}^{n} x_i$, then

Like a Binomial (several coin tosses) from several iid Bernoullis.

$$\begin{aligned} M_y(t) &= \mathcal{E}\left[e^{ty}\right] = \mathcal{E}\left[e^{tx_1}e^{tx_2}...e^{tx_n}\right] = \mathcal{E}\left[e^{tx_1}\right]\mathcal{E}\left[e^{tx_2}\right]...\mathcal{E}\left[e^{tx_n}\right] \\ &= M_{x_1}(t)M_{x_2}(t)...M_{x_n}(t) = \prod_{i=1}^{n} M_{x_i}(t) \end{aligned} \tag{1.14}$$

Chebyshev's theorem

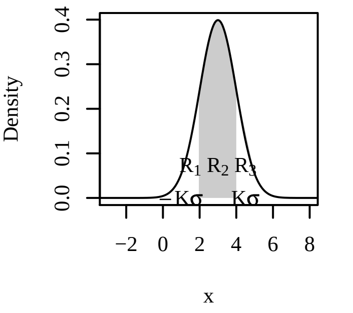


Fig. 1.6: Proof of Chebyshev's theorem.

If a probability distribution has mean μ and standard deviation σ, the probability of getting a value deviating from μ by at least $k\sigma$ is at most $1/k^2$. We can talk in terms of $k\sigma$, for example 2σ, 3σ etc.

$$P\left(\left|X - \mathcal{E}\left[X\right]\right| \geq k\sigma\right) \leq \frac{1}{k^2} \tag{1.15}$$

This holds for any probability distribution!

Proof:

$$\sigma_X^2 = \sum_{i=1}^{R} (x_i - \mu)^2 P(X = x_i)$$

Consider three regions R_1, R_2 and R_3 (see Fig. 1.6) where $R_1 : x \leq \mu - k\sigma$, $R_2 : \mu - k\sigma < x < \mu + k\sigma$ and $R_3 : x \geq \mu + k\sigma$. But $\left(x_i - \mathcal{E}\left[X\right]\right)^2 p(x_i) \geq 0$ always. Then

$$\sigma^2 = \sum_{R_1} \left(x_i - \mathcal{E}\left[X\right]\right)^2 p(x_i) + \sum_{R_2} \left(x_i - \mathcal{E}\left[X\right]\right)^2 p(x_i) + \sum_{R_3} \left(x_i - \mathcal{E}\left[X\right]\right)^2 p(x_i)$$

implies that

$$\sigma^2 \geq \sum_{R_1} (x_i - \mathcal{E}[X])^2 p(x_i) + \sum_{R_3} (x_i - \mathcal{E}[X])^2 p(x_i)$$

But in R_1, $x_1 - \mathcal{E}[X] \leq -k\sigma$ and in R_3, $x_i - \mathcal{E}[X] \geq k\sigma$. Hence in R_1 or R_3, $|x - \mathcal{E}[X]| \geq k\sigma$

$$\Rightarrow \sigma^2 \geq \sum_{R_1} k^2\sigma^2 p(x_i) + \sum_{R_3} k^2\sigma^2 p(x_i)$$

Chebyshev's theorem states that 75% of prob. mass is within 2σ of the mean. This is usually a poor upper limit. Approx 95 % of prob. mass falls within 2σ of the mean of most random variables. In fact, 95% of prob. mass lies within 1.96σ of the mean for a normal distribution.

This therefore implies that

$$\frac{1}{k^2} \geq \sum_{R_1} p(x_i) + \sum_{R_3} p(x_i)$$

$\sum_{R_1} p(x_i) + \sum_{R_3} p(x_i)$ is the probability of region $R_1 \cup R_3 = P(|X - \mathcal{E}[X]| \geq k\sigma)$.

Example 1.4. For 40000 tosses of a fair coin, there is at least a 0.99 prob that the proportion of heads will be between 0.475 and 0.525.

Soln:

To prove this, first note that for a binomial random variable, $\mathcal{E}[X] = 40000/2 = 20000 = np$. (The fact that $\mathcal{E}[X] = np$ will be proved later). Also $\sigma = \sqrt{np(1-p)} = \sqrt{40000 \times 1/2 \times 1/2} = 100$ and $1 - (1/k^2) = 0.99 \Rightarrow k = 10$. The probability is at least 0.99 that we get between $20000 - (10 \times 100) = 19000$ and $20000 + (10 \times 100) = 21000$ heads i.e. $19000/40000 = 0.475$ and $21000/40000 = 0.525$.

Transformations

Variance is the second moment of the deviation.

In general, given X = random variable, $P_X(r)$ = PMF of X, and $g(X)$ = some real valued function of X

$$\mathcal{E}[g(X)] = \sum_r g(r) P_X(r) \tag{1.16}$$

If $Y = aX + b$, then

$$\mathcal{E}[Y] = a\mathcal{E}[X] + b \tag{1.17}$$

and the variance is

b should not matter.

$$\begin{aligned}\text{Var}[Y] &= \sum_i (ax_i + b - \mathcal{E}[aX + b])^2 P_X(x_i) = \sum_i (ax_i + b - a\mathcal{E}[X] - b)^2 P_X(x_i) \\ &= \sum_i (a(x_i - \mathcal{E}[X]))^2 P_X(x_i) = a^2 \sum_i (x_i - \mathcal{E}[X])^2 P_X(x_i) = a^2 \text{Var}[X] \end{aligned} \tag{1.18}$$

Cumulative distribution function (cdf)

Notation: $F(X)$ = cdf for a specific value x

$$P(X \leq x) = F(X) \quad (\text{or } \Phi(x)) \tag{1.19}$$

The Bernoulli random variable

A Bernoulli random variable has two values 1 and 0. Example: coin toss: heads = p, tails = $1 - p$. $X = 1$ (heads) or 0 (tails). Then

The PMF can also be written as $p^x(1-p)^{1-x}$. Contrast this to the Binomial PMF below.

$$\text{PMF} = P_X(x_i) = \begin{cases} p & \text{if} \quad x_i = 1 \\ 1-p & \text{if} \quad x_i = 0 \end{cases} \tag{1.20}$$

This is a very simple 'on-off' distribution.

- $\mathcal{E}\left[X\right] = 1 \times p + 0 \times (1-p) = p$
- $\mathcal{E}\left[X^2\right] = 1^2 \times p + 0^2 \times (1-p) = p$ which implies that $\mathrm{Var}\left[X\right] = \mathcal{E}\left[X^2\right] - \left(\mathcal{E}\left[X\right]\right)^2 = p - p^2 = p(1-p)$, $\mathrm{Var}\left[X\right]$ is max when $p = 1/2$.

Discrete uniform random variable

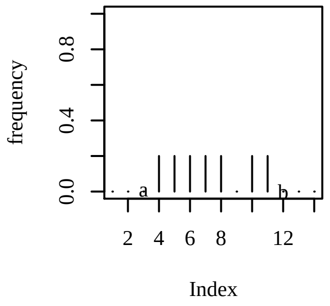


Fig. 1.7: Discrete Uniform random variable.

Example: The result of the roll of a die is a random variable $= X$.

- $\mathrm{PMF}(X) = P_X(x_i) = \begin{Bmatrix} 1/6 & \text{if} & i = 1,2,3,4,5,6 \\ 0 & & \text{otherwise} \end{Bmatrix}$
- $\mathcal{E}\left[X\right] = 3.5$ (using a symmetry argument about 3.5).
- $\mathrm{Var}\left[X\right] = 1/6 \times \left[1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right] - 3.5^2 = 35/12$.

The uniformly distributed random variable X takes one out of a range of integer values with equal probability (see Fig. 1.7). In general,

$$P_X(k) = \begin{cases} 1/(b-a+1) & \text{if} \quad k = a,\, a+1, ...b \\ 0 & \text{otherwise} \end{cases} \qquad a < b \tag{1.21}$$

The expectation is $\mathcal{E}\left[X\right] = (a+b)/2$. The variance of X is $(b-a)(b-a+2)/12$.

The Binomial Random variable

A biased coin ($H = p$, $T = 1-p$) is tossed n times. The toss outcomes are independent. $X =$ number of H in an n toss sequence = random var. (Fig. 1.8)

$$\mathrm{PMF}(X) = P_X(X = k) = {}^nC_k p^k (1-p)^{n-k}, \;\; k = 0,1,2,3.......n \tag{1.22}$$

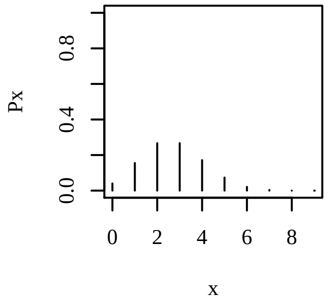


Fig. 1.8: A binomial distribution with $p = 0.5$ and $n = 9$.

- $\mathcal{E}\left[X\right] = \sum_{i=1}^{r} x_i P_X(X = x_i)$ r is from $0, 1, ..., n$. Using $q = 1-p$,

$$\begin{aligned}\mathcal{E}\left[X\right] &= \sum_{i=0}^{n} i\, {}^nC_i p^i q^{n-i} = \sum_{i=1}^{n} i \frac{n!}{i!(n-i)!} p^i q^{n-i} = np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} q^{n-i} \\ &= np[p + (1-p)]^{n-1} = np\end{aligned}$$

- $\mathcal{E}\left[X^2\right] = \sum_{i=1}^{r} x_i^2 P_X(X = x_i)$. Hence

$$\begin{aligned}\mathcal{E}\left[X^2\right] &= \sum_{i=0}^{n} i^2 \;\; {}^nC_i p^i q^{n-i} = \sum_{i=1}^{n} (i(i-1) + i) \;\; {}^nC_i p^i q^{n-i} \\ &= \sum_{i=1}^{n} \frac{i(i-1)n!}{(n-i)!i!} p^i (1-p)^{n-i} + \sum_{i=1}^{n} i\, {}^nC_i p^i (1-p)^{n-i} \\ &= n(n-1)p^2 \sum_{i=2}^{n} \text{‘} \frac{(n-2)!}{(i-2)!(n-i)!} p^{i-2} q^{n-i} + np \\ &= n(n-1)p^2 \left((p + (1-p))^2\right) + np = n(n-1)p^2 + np\end{aligned}$$

Therefore the variance is

$$\mathrm{Var}\left[X\right] = \mathcal{E}\left[X^2\right] - \left(\mathcal{E}\left[X\right]\right)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$$

Often q is used to denote $1-p$. Then $\mathrm{Var}\left[X\right] = npq$.

$\mathcal{E}\left[X\right] = np \Rightarrow$ Success in n trials $= n$ times success in one trial $\Rightarrow$ intuitive relationship to Bernoulli. The max of $\mathrm{Var}\left[X\right]$ is at $p = 1/2$. $\mathrm{Var}\left[X\right] \to 0$ as $p \to 0$ or $p \to 1$, which shows a skew towards extremes when $p \to 0$ or $p \to 1$. This implies that there is clustering around that extreme and very little variability.

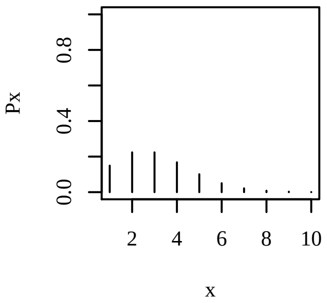


Fig. 1.9: The Poisson distribution with $\lambda = 3$.

The Poisson random variable

The Poisson variable is usually associated with rare events (births, deaths, accidents, radioactivity, mutations, phone calls etc.). Let the average number of calls per day be known. Then the probability of getting a call in the next hour ($\Delta t = 1$ hour) is proportional to $\Delta t = \lambda \Delta t$ using some constant λ. The probability of not getting a call in $\Delta t = 1 - \lambda \Delta t$. In this Δt, it is very unlikely that we get two calls. The average call rate does not change with time. Similarly, if the average birth rate can be calculated over many months, then the probability of observing a birth depends on the time interval Δt. The birth rate could vary depending on population size, hence it may not be a Poisson distribution over a very long time. A major assumption made is that what happens in one time interval is independent of the previous interval. (Fig. 1.9).

$$\text{PMF}(X) = P_X(k) = e^{-\lambda}(\lambda^k/k!),\ k = 0, 1, 2... \tag{1.23}$$

- Is P_X really a PMF?

$$\sum_{k=0}^{\infty} P_X(k) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...\right) = e^{-\lambda}e^{\lambda} = 1$$

- $\mathcal{E}[X]$ is (using $m = k - 1$),

$$\mathcal{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \lambda \sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} = \lambda$$

- $\text{Var}[X] = \mathcal{E}\left[X^2\right] - (\mathcal{E}[X])^2 = \lambda$

$$\begin{aligned}\mathcal{E}\left[X^2\right] &= \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k(k-1) e^{-\lambda}\frac{\lambda^k}{k!} + \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda^2 \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda\end{aligned}$$

$$\Rightarrow \text{Var}[X] = \mathcal{E}\left[X^2\right] - (\mathcal{E}[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

For a Poisson distribution, mean = variance = λ.

Given a data set with mean $\simeq$ variance, it may be described by a Poisson.

Poisson vs Binomial

For a Binomial: mean = np, variance = npq. For a Poisson variable, mean = variance. Hence when $p \simeq 0$, $q =\simeq 1$, then $np \simeq npq$. The Poisson is like a Binomial with very small p, very large n. If $\lambda = np$, with n= large and p is small, then

$$e^{-\lambda}\frac{\lambda^k}{k!} \simeq {}^nC_k p^k (1-p)^{n-k}$$

Proof Using the binomial equation $p(X) = {}^nC_x p^x (1-p)^{n-x}$, and requiring that $\lambda = np \Rightarrow p = \lambda/n$ which gives

$$\begin{aligned}p(x) &= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{n(n-1)...(n-x+1)}{x!n^x}\lambda^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{(1-1/n)(1-2/n)...(1-(x-1)/n)}{x!}\lambda^x \left(1 - \frac{\lambda}{n}\right)^{n-x}\end{aligned}$$

as $n \to \infty$, $(1-1/n)(1-2/n)...(1-(x-1)/n) \to 1$ and

$$\left(1 - \frac{\lambda}{n}\right)^{n-x} = \left[\left(1 - \frac{\lambda}{n}\right)^{n/\lambda}\right]^{\lambda} \left(1 - \frac{\lambda}{n}\right)^{-x} \to e^{-\lambda}$$

and therefore

$$p(X) \simeq e^{-\lambda}\frac{\lambda^k}{k!}, \quad x = 0, 1, 2...$$

□

Use Poisson instead of Binomial for $n \geq 20$ and $p \leq 0.05$. $\lambda = np$.

If feasible use Poisson instead of Binomial [you would calculate ${}^nC_k$ and $(1-p)^{n-k}$ for a Binomial, and this is difficult for large n]. This approximation is valid/reasonable when $n \geq 100$, $p \leq 0.01$ and at slightly higher error for $n \geq 20$ and $p \leq 0.05$.

The geometric random variable

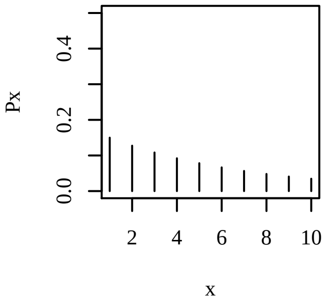


Fig. 1.10: The Geometric distribution with $p = 0.15$.

Consider a biased coin ($H = p$, $T = 1-p$, $0 < p < 1$). Let X = random variable = # of tosses when H comes up for the first time. Then the PMF of X is (Fig. 1.10).

$$P_X(k) = (1-p)^{k-1}p, \; k = 1, 2... \tag{1.24}$$

- Proof that this is a PMF:

$$\sum_{k=1}^{\infty} P_X(k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = p\sum_{k=0}^{\imath nfty}(1-p)^k = p\frac{1}{1-(1-p)} = 1$$

  Instead of the first H, we could look for the number of repeated independent trials till first success. We have a geometric progression with parameter $(1-p)$.

- The mean and variance

$$\mathcal{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p, \qquad \text{Var}[X] = \sum_{k=1}^{\infty}(k - \mathcal{E}[X])^2(1-p)^{k-1}p$$

The mean of a geometric random variable is $1/p$ and the variance is $(1-p)/p^2$.

Evaluating these sums are difficult. Instead, we make use of a neat trick. Let $A_1 = \{X = 1\}$ and $A_2 = \{X > 1\}$. Then if $X = 1$, $\mathcal{E}[X|X=1] = 1$ (we have success on the first try). If the first try fails ($X > 1$), one try is wasted, but we are back where we started ⇒ expected number of remaining tries is still $\mathcal{E}[X]$ and therefore $\mathcal{E}[X|X>1] = 1 + \mathcal{E}[X]$

$$\begin{aligned}\mathcal{E}[X] &= P(X=1)\mathcal{E}[X|X=1] + P(X>1)\mathcal{E}[X|X>1]\\ &= p + (1-p)(1+\mathcal{E}[X]) = \frac{1}{p} \text{ (on rearranging)}\end{aligned}$$

Similarly $\mathcal{E}\left[X^2|X=1\right] = 1$ which implies that

$$\mathcal{E}\left[X^2|X>1\right] = \mathcal{E}\left[(1+\mathcal{E}[X])^2\right] = 1 + 2\mathcal{E}[X] + \mathcal{E}\left[X^2\right]$$

Therefore,

$$\begin{aligned}\mathcal{E}\left[X^2\right] &= p \times 1 - (1-p)(1 + 2\mathcal{E}[X] + \mathcal{E}\left[X^2\right])\\ &= \frac{1 + 2(1-p)\mathcal{E}[X]}{p} = \frac{2}{p^2} - \frac{1}{p}\\ \text{Var}[X] = \mathcal{E}\left[X^2\right] - (\mathcal{E}[X])^2 &= \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}\end{aligned}$$

Multiple random variables: Joint PMFS

Given two discrete random variables X and Y in the same experiment, the joint PMF of X and Y is $P_{X,Y}(X = x, Y = y)$ for all pairs of (x, y) that X and Y can take.

$$P_X(x) = \sum_y P_{X,Y}(x,y) \qquad\qquad P_Y(y) = \sum_x P_{X,Y}(x,y) \tag{1.25}$$

$P_X(x)$ and $P_Y(y)$ are called marginal PMFs. If $Z = g(X,Y)$, $P_Z(z) = \sum P_{X,Y}(x,y)$ where the summation is over all (x, y).

$$P_Z(z) = \sum_{\{(x,y)|g(x,y)=z\}} P_{X,Y}(x,y) \tag{1.26}$$

$$\mathcal{E}\left[g(x,y)\right] = \sum_{x,y} g(x,y)P_{X,Y}(x,y) \tag{1.27}$$

For example, $\mathcal{E}\left[aX + bY + c\right] = a\mathcal{E}\left[X\right] + b\mathcal{E}\left[Y\right] + c$.

For three variables:

$$P_{X,Y}(x,y) = \sum_{z} P_{X,Y.Z}(x,y,z) \tag{1.28}$$

$$P_X(x) = \sum_{y}\sum_{z} P_{X,Y.Z}(x,y,z) \tag{1.29}$$

$$\mathcal{E}\left[g(X,Y,Z)\right] = \sum_{x,y,z} g(x,y,z)P_{X,Y.Z}(x,y,z) \tag{1.30}$$

For n random variables $X_1, X_2...X_n$ and scalars $a_1, a_2...a_n$

$$\mathcal{E}\left[a_1X_1 + a_2X_2 + ... + a_nX_n\right] = a_1\mathcal{E}\left[X_1\right] + a_2\mathcal{E}\left[X_2\right] + ... + a_n\mathcal{E}\left[X_n\right]$$

We saw this before: a binomial is constructed from several Bernoulli variables.

For example,

$$\mathcal{E}\left[aX + bY + cZ + d\right] = a\mathcal{E}\left[X\right] + b\mathcal{E}\left[Y\right] + c\mathcal{E}\left[Z\right] + d$$

Conditional PMFs

Given that event A has occurred, we can use conditional probability: this implies a new sample space where A is known to have happened. The conditional PMF of X, given that event A occurred with $P(A) > 0$ is

$$P_{X|A}(x) = P(X = x|A) = \frac{P(\{X = x\} \cap A)}{P(A)} \tag{1.31}$$

$$\sum_{x} P(\{X = x\} \cap A) = P(A) \tag{1.32}$$

Events $\{X = x\} \cap A$ are disjoint for different values of x and hence $\sum P_{X|A} = 1 \Rightarrow P_{X|A}$ is a PMF.

Example 1.5. X = roll of a die and A = event that roll is an even number

$$P_{X|A} = P(X = x|\text{roll is even})\frac{P(X = x \text{ and } X \text{ is even})}{P(\text{roll is even})} = \begin{cases} 1/3 & \text{if } x = 2,4,6 \\ 0 & \text{otherwise} \end{cases}$$

Conditional PMF of X, given random variable $Y = y$ is used when one random variable is conditioned on another.

$$P_{X|Y} = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$

Note that $\sum_x P_{X|Y}(x|y) = 1$ for normalization. $P_{X,Y}(x,y) = P_Y(y)P_{X|Y}(x|y) = P_X(x)P_{Y|X}(y|x) \Rightarrow$ we can compute joint PMFs from conditionals using a sequential approach.

Example 1.6. Consider four rolls of a six sided die. X = number of 1's and Y = number of 2's obtained. What is $P_{X,Y}(x,y)$?

Soln:

The marginal PMF P_Y is a binomial.

$$P_Y(y) = {}^4C_y(1/6)^y(5/6)^{4-y}, \quad y = 0, 1, 2, 3, 4$$

To get conditional PMFs $P_{X|Y}(x|y)$ given that $Y = y$, $X =$ number of 1's in the $4 - y$ rolls remaining, $\Rightarrow$ values $1, 3, 4, 5, 6$ occur with equal probability. Therefore

$$P_{Y|X}(y|x) =^{4-y} C_x(1/5)^x(4/5)^{4-y-x} \text{ for all } x, y = 0, 1, 2, 3, 4 \text{ and } 0 \le x + y \le 4$$

For x, y such that $0 \le x + y \le 4$

$$\begin{aligned} P_{X,Y}(x, y) &= P_Y(y)P_{X|Y}(x|y) \\ &= \left({}^4C_y(1/6)^y(5/6)^{4-y}\right) \times \left({}^{4-y}C_x(1/5)^x(4/5)^{4-y-x}\right) \end{aligned}$$

For other (x, y), $P_{X,Y}(x, y) = 0$

This is the total probability theorem in a different notation.

Marginal PMFs can therefore be computed from conditional PMFs:

$$P_X(x) = \sum_y P_{X,Y}(x, y) = \sum_y P_Y(Y = y)P_{X|Y}(x|y)$$

Independence

Random variable X is independent of event A if $P(X = x \,\text{and}\, A) = P(X = x)P(A) = P_X(x)P(A)$ for all x. As long as $P(A) > 0$, since $P_{X|A}(x) = P(X = x \,\text{and}\, A)/P(A)$ this implies that $P_{X|A}(x) = P_X(x)$ for all x.

The value of $Y(= y)$ has no bearing on the value of $X(= x)$.

Random variables X and Y are independent if

$$P_{X,Y}(x, y) = P_X(x)P_Y(y) \tag{1.33}$$

for all $x, y \Rightarrow P_{X|Y}(x, y) = P_X(x)$ for all y with $P_Y(y) > 0$, and all x.

$p(A) > 0$.

Random variables X and Y are conditionally independent given an event A if

$$P(X = x, Y = y|A) = P(X = x|A)P(Y = y|A) \tag{1.34}$$

for all x and y . $P_{X|Y}(x|y) = P_{X|A}(x)$ for all x and y such that $P_{Y|A}(y) > 0$.

Independence implies that $\mathcal{E}[XY] = \mathcal{E}[X]\,\mathcal{E}[Y]$.

For independent variables,

$$\begin{aligned} \mathcal{E}[XY] &= \sum_x\sum_y xyP_{X,Y}(x, y) = \sum_x\sum_y xyP_X(x)P_Y(y) \\ &= \sum_x xP_X(x)\sum_y yP_Y(y) = \mathcal{E}[X]\,\mathcal{E}[Y] \end{aligned} \tag{1.35}$$

Similarly

$$\mathcal{E}[g(X)h(Y)] = \mathcal{E}[g(X)]\,\mathcal{E}[h(Y)] \tag{1.36}$$

Expectations of combined variables always add up. Variances add up only if X and Y are independent.

For two independent random variables X and Y,

$$\mathcal{E}[X + Y] = \mathcal{E}[X] + \mathcal{E}[Y] \tag{1.37}$$

$$\begin{aligned} \mathrm{Var}[Z] &= \mathcal{E}\left[(X + Y - \mathcal{E}[X + Y])^2\right] = \mathcal{E}\left[(X + Y - \mathcal{E}[X] - \mathcal{E}[Y])^2\right] \\ &= \mathcal{E}\left[((X - \mathcal{E}[X]) + (Y - \mathcal{E}[Y]))^2\right] \\ &= \mathcal{E}\left[(X - \mathcal{E}[X])^2\right] + \mathcal{E}\left[(Y - \mathcal{E}[Y])^2\right] + 2\mathcal{E}[(X - \mathcal{E}[X])](Y - \mathcal{E}[Y])] \\ &= \mathcal{E}\left[(X - \mathcal{E}[X])^2\right] + \mathcal{E}\left[(Y - \mathcal{E}[Y])^2\right] = \mathrm{Var}[X] + \mathrm{Var}[Y] \end{aligned} \tag{1.38}$$

Independence for several random variables

Independence $\Rightarrow P_{X,Y,Z}(x,y,z) = P_X(x)P_Y(y)P_Z(z)$ for all x, y, z etc. Obviously $g(X,Y)$and $h(Z)$ are independent. For n independent variables $X_1, X_2...X_n$

$$\mathrm{Var}\left[X_1 + X_2 + ...X_n\right] = \mathrm{Var}\left[X_1\right] + \mathrm{Var}\left[X_2\right] + ... + \mathrm{Var}\left[X_n\right] \tag{1.39}$$

Example 1.7. Variance of the binomial: Given n independent coin tosses , with $p =$ probability of heads. For each i let $X_i =$ Bernoulli random variable ($= 1$ if heads, $= 0$ if tails) $\Rightarrow X = X_1 + X_2 + ...X_n$. Then X is a binomial random variable. Coin tosses are independent $\Rightarrow X_1, X_2....X_n$ are independent variables and

$$\mathrm{Var}\left[X\right] = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = np(1-p)$$

Covariance and Correlation

Covariance of X & Y (random vars) = $\mathrm{cov}\left[X, Y\right]$

$$\mathrm{cov}\left[X,Y\right] = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])(Y - \mathcal{E}\left[Y\right])\right] = \mathcal{E}\left[XY\right] - \mathcal{E}\left[X\right]\mathcal{E}\left[Y\right] \tag{1.40}$$

When $\mathrm{cov}\left[X,Y\right] = 0$, X & Y are uncorrelated. If X & Y are independent,

$$\mathcal{E}\left[XY\right] = \mathcal{E}\left[X\right]\mathcal{E}\left[Y\right] \qquad \Rightarrow \qquad \mathrm{cov}\left[X,Y\right] = 0$$

If X & Y are independent, they are also uncorrelated.

Example 1.8. Let (X, Y) take values $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$ each with probability $1/4$ i.e. marginal probability of X & Y are symmetric around 0. So $\mathcal{E}\left[X\right] = \mathcal{E}\left[Y\right] = 0$. For all possible values of (x, y), either x or $y = 0$.

$$\Rightarrow XY = 0 \Rightarrow \mathcal{E}\left[XY\right] = 0 \Rightarrow \mathrm{cov}\left[X,Y\right] = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])(Y - \mathcal{E}\left[X\right])\right] = \mathcal{E}\left[XY\right] = 0$$

$\Rightarrow X$ & Y are uncorrelated. However X & Y are not independent: nonzero X implies $Y = 0$ etc.

The reverse is not true! Beware!

Uncorrelated variables need not be independent!

Correlation Coefficient of X & Y

Let X & Y = random variables with nonzero variances. Correlation coefficient = normalized version of covariance =

$$\rho_{xy} = \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}} \tag{1.41}$$

Note that $-1 \leq \rho_{XY} \leq 1$.

Example 1.9. Consider n independent tosses of a biased coin (head $= p$). Let $X = \#$ of heads & $Y = \#$ of tails. Are X, Y correlated?

Soln:

For all (x, y), $x + y = n \Rightarrow \mathcal{E}\left[X\right] + \mathcal{E}\left[Y\right] = n \Rightarrow X - \mathcal{E}\left[X\right] = -(Y - \mathcal{E}\left[Y\right])$

$$\mathrm{cov}\left[X,Y\right] = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])(Y - \mathcal{E}\left[X\right])\right] = -\mathcal{E}\left[(X - \mathcal{E}\left[X\right])^2\right] = -\mathrm{Var}\left[X\right]$$

Note that $\mathrm{Var}\left[Y\right] = \mathrm{Var}\left[n - X\right] = \mathrm{Var}\left[-X\right] = \mathrm{Var}\left[X\right]$.

$$\Rightarrow \rho(X,Y) = \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}} = \frac{-\mathrm{Var}\left[X\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}} = -1$$

Correlation

For the sum of several, not necessarily independent variables

General case.

$$\mathrm{Var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left[X_i\right] + 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \mathrm{cov}\left[X_i X_j\right], \quad i < j \tag{1.42}$$

$$\begin{aligned}\mathrm{Var}\left[c_1 X_1 + c_2 X_2\right] &= c_1^2 \mathrm{Var}\left[X_1\right] + c_2^2 \mathrm{Var}\left[X_2\right] + 2c_1 c_2 \mathrm{cov}\left[X_1 X_2\right] \\ &= c_1^2 \mathrm{Var}\left[X_1\right] + c_2^2 \mathrm{Var}\left[X_2\right] \\ &\quad +2c_1 c_2 \sqrt{\mathrm{Var}\left[X_1\right] \mathrm{Var}\left[X_2\right]} \rho(X_1, X_2)\end{aligned}$$

1.3 Continuous probability distributions

Continuous vs Discrete: Continuous distribs are more fine grained, may be more accurate and permit analysis by calculus. A random variable X is continuous if its probability distribution can be defined by a non negative function f_X of X such that

$$P(X) = \int f_X(x)dx \tag{1.43}$$

i.e P(X is between a and b) $= \int_a^b f_X(x)dx =$ area under the curve from a to b. Total area under the curve $= 1$. $f_X(x)$ is a probability density function (PDF). Intuitively, PDF of X is large for high probability and low for low probability. For a single value of X, $P(X = a) = \int_a^a f_X(x)dx = 0$. So including endpoints does not matter

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

For f_X to be a valid PDF, it must be

$f_X(x)$ is not the probability of an event: it can even be greater than one.

1 Positive for all x, $f_X(x) \geq 0$ for all x
2 $\int_{-\infty}^{\infty} f_X(x)dx = P(-\infty < X < \infty) = 1$

$P([x, x + \triangle x]) = \int_x^{x+\triangle x} f_X(t)dt \approx f_X(x)\triangle x$ where $f_X(x) =$ probability mass per unit length near x.

Expectation and Moments for continuous RVs

$$\mathcal{E}\left[X\right] = \int_{-\infty}^{\infty} x f_X(x)dx \tag{1.44}$$

The n^{th}moment is $\mathcal{E}\left[X^n\right]$. The variance is

$\mathcal{E}\left[g(x)\right] = \int_{-\infty}^{\infty} g(x) f_X(x)dx$, $g(x)$ may be continuous or discrete.

$$\mathrm{Var}\left[X\right] = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])^2\right] = \mathcal{E}\left[X^2\right] - (\mathcal{E}\left[X\right])^2 = \int_{-\infty}^{\infty} (x - \mathcal{E}\left[X\right])^2 f_X(x)dx > 0$$

If $Y = aX + b$, $\mathcal{E}\left[Y\right] = a\mathcal{E}\left[X\right] + b$ and $\mathrm{Var}\left[Y\right] = a^2 \mathrm{Var}\left[X\right]$

Continuous uniform random variable

Discrete Uniform law: Sample space had finite number of equally likely outcomes. For discrete variables, we count the number of outcomes concerned with an event. For continuous variables we compute the length of a subset of the real line.

The PDF of a continuous uniform random variable is

$$f_X(x) = \begin{cases} c, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{1.45}$$

where c must be > 0. For f to be a PDF, $1 = \int_a^b c\,dz = c\int_a^b dz = c(b - a) \Rightarrow c = 1/(b - a)$

Piecewise constant PDF, general form:

$$f_x(x) = \begin{cases} c_i & \text{if } a_i \le x \le a_{i+1},\ \ i = 1, 2, ..., n-1 \\ 0 & \text{otherwise} \end{cases}$$

$$1 = \int_{a_1}^{a_n} f_X(x)dx = \sum_{i=1}^{n-1} \int_{a_i}^{a_n} c_i dx = \sum_{i=1}^{n-1} c_i(a_{i-1} - a_i)$$

$$f_X(x) = \begin{cases} 1/2\sqrt{x} & \text{if } 0 < x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{-\infty}^{\infty} f_X(x) = \int_0^1 \frac{1}{2\sqrt{x}}dx = \sqrt{x}\Big|_0^1 = 1$$

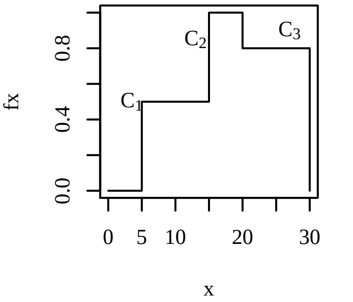


Fig. 1.11: The piecewise constant distribution.

Mean and variance of the uniform random variable:

$$\mathcal{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{b-a} \times \frac{1}{2} \times x^2\Big|_a^b$$
$$= \frac{1}{b-a} \times \frac{b^2 - a^2}{2} = \frac{a+b}{2}$$

The PDF is symmetric around $(a+b)/2)$

$$\mathcal{E}\left[X^2\right] = \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{b-a} \times \frac{x^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$
$$\text{Var}[X] = \mathcal{E}\left[X^2\right] - (\mathcal{E}[X])^2 = \frac{(a^2 + ab + b^2)}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Example 1.10. Let $[a, b] = [0, 1]$ and let $g(x) = \begin{cases} 1 & \text{if } x \le 1/3 \\ 2 & \text{if } x > 1/3 \end{cases}$

1. Discrete: $Y = g(X)$ has PMF $\begin{cases} P_Y(1) = P(X \le 1/3) = 1/3 \\ P_Y(2) = P(X > 1/3) = 2/3 \end{cases}$

$$\mathcal{E}[Y] = \frac{1}{3} \times 1 + \frac{2}{3} \times 2 = \frac{5}{3}$$

2. Continuous:

$$\mathcal{E}[Y] = \int_0^1 g(x) f_x(x)dx = \int_0^{1/3} 1dx + \int_{1/3}^{2/3} 2dx = \frac{5}{3}$$

Exponential continuous distribution

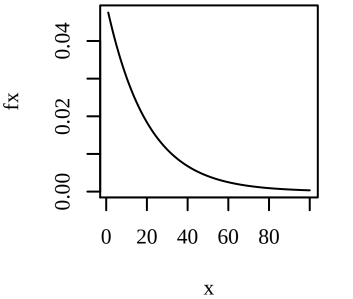


Fig. 1.12: The exponential distribution with $\lambda = 0.05$.

PDF of X is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.46}$$

- Is it a PDF?

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^{\infty} = 1$$

For any $a \ge 0$

$$P(X \ge a) = \int_a^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_a^{\infty} = e^{-\lambda a}$$

Examples: Time till a light bulb burns out. Time till an accident.

- Mean

$$\mathcal{E}\left[X\right] = \int_0^\infty x(\lambda e^{-\lambda x})dx = (-xe^{-\lambda x})|_0^\infty + \int_0^\infty e^{-\lambda x}dx = 0 - \frac{e^{-\lambda x}|_0^\infty}{\lambda} = \frac{1}{\lambda}$$

- Variance

$$\mathcal{E}\left[X^2\right] = \int_0^\infty x^2(\lambda e^{-\lambda x})dx = (-x^2e^{-\lambda x})|_0^\infty + \int_0^\infty 2xe^{-\lambda x}dx = 0 + \frac{2}{\lambda}\mathcal{E}\left[X\right] = \frac{2}{\lambda^2}$$

$$\mathrm{Var}\left[X\right] = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Example 1.11. The sky is falling... Meteorites land in a certain area at an average of 10 days. What is the probability that the first meteorite lands between 6 AM and 6 PM of the first day given that it is now midnight.

Soln:

Let X = elapsed time till strike (in days) = exponential variable $\Rightarrow$ mean $= 1/\lambda = 10$ days $\Rightarrow \lambda = 1/10$.

$$P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = P\left(X \geq \frac{1}{4}\right) - P\left(X \geq \frac{3}{4}\right)$$
$$= e^{-\lambda(1/4)} - e^{-\lambda(3/4)} = e^{-1/40} - e^{-3/40} = 0.047$$

Probability that meteorite lands between 6 AM and 6 PM of some day:
For k^{th} day, this time frame is $(k - 3/4) \leq X \leq (k - 1/4)$

$$= \sum_{k=1}^{\infty} P\left(\left(k - \frac{3}{4}\right) \leq X \leq \left(k - \frac{1}{4}\right)\right) = \sum_{k=1}^{\infty} P\left(X \geq \left(k - \frac{3}{4}\right)\right) - \sum_{k=1}^{\infty} P\left(X \geq k - \frac{1}{4}\right)$$
$$= \sum_{k=1}^{\infty} (e^{-(4k-3)/40} - e^{-(4k-1)/40})$$

Cumulative distribution function (CDF)

CDF of X

$$F_X = P(X \leq x) = \begin{cases} \sum_{k \leq x} P_X(k) & X: \text{Discrete} \\ \int_{-\infty}^{x} f_X(t)dt & X: \text{Continuous} \end{cases} \tag{1.47}$$

For the discrete form the CDF will look like a staircase.

Some properties of CDFs $F_X(x)$

- $F_X(x)$ is monotonically nondecreasing: if $x \leq y$, then $F_X(x) \leq F_X(y)$.
- $F_X(x) \to 0$ as $x \to -\infty$, and $F_X(x) \to 1$ as $x \to +\infty$.
- If X is discrete, F_X is piecewise constant ('staircase').
- If X is continuous, F_X is continuous.
- If X is discrete:
  - CDF from PMF:

$$F_X(k) = \sum_{i=-\infty}^{k} P_X(i)$$

  - PMF from CDF: (for all k)

$$P_X(k) = P(X \leq k) - P(X \leq k-1) = F_X(k) - F_X(k-1)$$

- If X is continuous:

– CDF from PMF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

– PMF from CDF:

$$f_X(x) = \left.\frac{dF_X}{dx}\right|_x$$

Geometric vs. Exponential CDF's

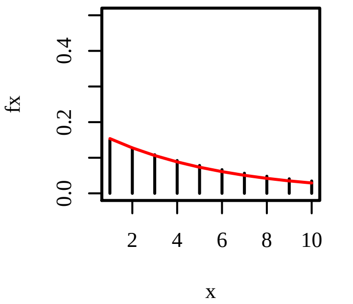


Fig. 1.13: Geometric vs. exponential distributions ($p = 0.15$ and $\lambda = 0.185$).

Let X = geometric random var. with parameter p. For example, X = number of trials before first success with probability of success per trial = p.

$$P(X = k) = p(1-p)^{k-1},\ for\, k = 1, 2, ...$$

$$F_X^{geo}(n) = \sum_{k=1}^{n} p(1-p)^{k-1} = p \times \frac{1-(1-p)^n}{1-(1-p)} = 1-(1-p)^n, \quad n = 1, 2...$$

Let Y = exponential random var, $\lambda > 0$

$$\Rightarrow F_Y^{exp}(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t}|_0^x = 1 - e^{-\lambda x}, \quad \text{for}\, x > 0$$

Comparison: Let

$$\delta = -\frac{\ln(1-p)}{\lambda} \Rightarrow e^{-\lambda\delta} = 1 - p$$

$\Rightarrow$ Exponential and geometric CDF's match for all $x = n\delta$, $n = 1, 2, ...$

$$\Rightarrow F^{exp}(n\delta) = F^{geo}(n),\ n = 1, 2...$$

Normal (Gaussian) random variables

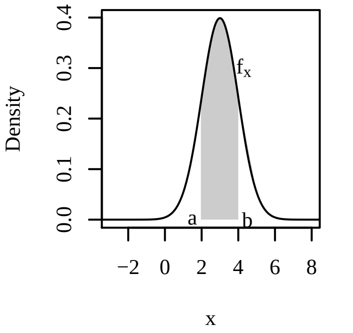


Fig. 1.14: Normal distribution.

Most useful distribution: use it to approximate other distributions

$$\text{PDF of } X = f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty \tag{1.48}$$

for some parameters μ & σ ($\sigma > 0$).

- Normalization confirms that it is a PDF:

$$\int_{-\infty}^{\infty} f_X(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1$$

- $\mathcal{E}[X]$ = mean = μ (from symmetry arguments alone).
- $\text{Var}[X] =$

$$\int_{-\infty}^{\infty} (x - \mathcal{E}[X])^2 f_X(x)dx = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

Let $z = (x-\mu)/\sigma$

$$\begin{aligned}\text{Var}[x] &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{\sigma^2}{\sqrt{2\pi}} \times \left(-ze^{-z^2/2}\right)\Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = \sigma^2\end{aligned}$$

X = normal var with μ and σ^2= (mean,var). Height of $\mathcal{N}(\mu, \sigma) = 1/\sigma\sqrt{2\pi}$ and hence height $\propto 1/\sigma$. If $Y = aX + b$ then $\mathcal{E}[Y] = a\mathcal{E}[X] + b$ and $\text{Var}[Y] = a^2\sigma^2$ where X and Y are independent.

Standard Normal Distribution

A normal distribution of Y with mean $= 0$ and $\sigma^2 = 1$; $\mathcal{N}(\mu,\sigma^2) = \mathcal{N}(0,1) =$ standard normal

$$\text{CDF of } Y = \Phi(y) = P(Y \le y) = P(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt \tag{1.49}$$

$\Phi(y)$ is available as a table. If the table only gives the value of $\Phi(y)$ for $y \ge 0$, use symmetry.

Example 1.12. $\Phi(-0.5) = P(Y \le -0.5) = P(Y \ge 0.5) = 1 - P(Y \le 0.5) = 1 - \Phi(0.5) = 1 - 0.6915 = 0.3085$

Given a normal random var X with mean μ and variance σ^2, use

$$Z = \frac{(x-\mu)}{\sigma}$$

Then

$$\mathcal{E}[Z] = \frac{(\mathcal{E}[X] - \mu)}{\sigma} = 0$$

$$\text{Var}[Z] = \frac{\text{Var}[X]}{\sigma^2} = 1$$

$\Rightarrow Z =$ standard normal.

Example 1.13. Average rain at a spot = normal distribution, mean $= \mu = 60$ mm/yr and σ= 20 mm/yr. What is the probability that we get at least 80 mm/yr?

Soln:

Let X = rainfall/yr. Then $Z = X - \mu/\sigma = (X - 60)/20$. So

$$P(X \ge 80) = P(Z \ge \frac{80-60}{20}) = P(Z \ge 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

In general

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

where Z is standard normal. $100 \times u^{th}$ percentile of $\mathcal{N}(0,1) = Z_u$ and $P(X < Z_u) = u$ where $X \sim N(0,1)$

Approximating a binomial with a normal

Binomial: Probability of k successes in n trials, each trial has success probability of p. The binomial PMF is painful to handle for large n, so we approximate. $E[X] = np$ and $\text{Var}[X] = npq \Rightarrow$ the equivalent Normal distribution is $\mathcal{N}(np, npq)$.

Consider a binomial with mean $= np$ and variance $= npq$, with a range of interest from $a \le X \le b$. It may be approximated with a normal, $\mathcal{N}(np, npq)$ from $a - 1/2 \le X \le b + 1/2$. For the special case: $a = b$, $P(X = a) =$ area under normal curve from $a - 1/2$ to $a + 1/2$.

$$\text{At the extremes:} \begin{cases} \text{Binomial} & \text{Normal} \\ P(X=0) & \text{area to the left of } 1/2 \\ P(X=n) & \text{area to the right of } n - 1/2 \end{cases}$$

Use when $npq \ge 5$, n, p, q all moderate.

Approximating the Poisson with a normal

Poisson is difficult to use for large μ. $P_X(X = x) \simeq \mathcal{N}(\lambda, \lambda)$ where for the Poisson, mean = var = λ. For the Normal, we take the area from $x - 1/2$ to $x + 1/2$ or the area to left of $1/2$ for $x = 0$. Use this approximation for $\lambda \ge 10$.

Summary of approximations

1. Approximating Binomial with Normal:
   Rule of Thumb: $npq \ge 5$, n, p, q all moderately valued.
2. Approximating Poisson with Normal: $\lambda \ge 10$
3. Approximating Binomial with Poisson: $n \ge 100$, $p \le 0.01$
   (Obviously, both Poisson and Normal may be used for $n \ge 1000$, $p = 0.01$)

Beta distribution

Unlike the binomial, a and b need not be integers.

When $a = b = 1$, Beta(x;1,1) = Uniform distribution.

The Beta distribution has as its pdf

$$\text{PMF}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \qquad 0 \leq x \leq 1 \tag{1.50}$$

$\Gamma(a+1) = a\Gamma(a)$ in general and when a is an integer, using this relationship recursively gives us $\Gamma(a+1) = a!$. It is important to appreciate that by varying a and b, the pdf could attain various shapes. This is particularly useful when we look for a pdf of an appropriate shape.

$$\begin{aligned}
\mathcal{E}\left[x\right] &= \int_0^1 x f(x) dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a (1-x)^{b-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^{b-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}
\end{aligned}$$

$$\mathcal{E}\left[x^2\right] = \frac{a(a+1)}{(a+b+1)(a+b)}$$

Similarly

$$\text{Var}\left[x\right] = \frac{ab}{(a+b)^2(a+b+1)}$$

2 Estimation and Hypothesis Testing

2.1 Samples and statistics

There's no sense in being precise when you don't even know what you're talking about. *John von Neumann*

- We sample to draw some conclusions about an entire population.
  Examples: Exit polls, census.
- We assume that the population can be described by a probability distribution.
- We assume that the samples give us measurements following this distribution,
- We assume that the samples are randomly chosen. It is very difficult to verify that a sample is really a random sample. The population is assumed to be (effectively) infinite.
- Sampling theory must be used if the population is actually finite.
- Simple random sample: each member of a random sample is chosen independently and has the same (nonzero) probability of being chosen.
- Cluster sampling: If the entire population cannot be uniformly sampled, divide it into clusters and sample some clusters.

Clinical Trials and Randomization

- Block randomization: one block gets the treatment and another the placebo.
- Blind trials: Patients do not know what they get.
- Double blind trials: both doctors and patients do not know what they get.

Distributions and Estimation

- We wish to infer the parameters of a population distribution.
  - mean, variance, proportion etc.
- We use a statistic derived from the sample data.
  - Sample mean, sample variance etc.
- A statistic is a random variable.
  - What distribution does it follow?
  - What are its parameters?

Notation

- Population parameters are typically in Greek.
  - μ = population mean.
  - σ^2 = population variance.
- The corresponding statistic is not in Greek!
  - $\bar{x}$ = sample mean.
  - s^2 = sample variance.
- We can also denote an estimate of the parameter θ as $\hat{\theta}$.
- Regression:
  - You want: $y = \alpha x + \beta$.
  - You get: $y = ax + b$.

Why do we infer μ using the sample mean?

Consider a population with mean μ and variance σ^2. We have n measurements, x_1, x_2, …, x_n.

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} \tag{2.1}$$

It's all about expectations

$$\mathcal{E}[x_i] = \mu \qquad \text{Var}[x_i] = \sigma^2 \tag{2.2}$$

$$\mathcal{E}[\bar{x}] = \mathcal{E}\left[\frac{x_1 + \cdots + x_n}{n}\right] = \frac{\mathcal{E}[x_1] + \cdots + \mathcal{E}[x_n]}{n} = \frac{n\mu}{n} = \mu \tag{2.3}$$

$$\mathcal{E}[x_i] = \mu \qquad \mathcal{E}[\bar{x}] = \mu \tag{2.4}$$

$\bar{x}$ is an unbiased estimator of μ. For a statistic $\hat{\theta}$ to be unbiased

$$\mathcal{E}\left[\hat{\theta}\right] = \theta \tag{2.5}$$

The population mean has other unbiased estimators such as the sample median, (min + max)/2 and the trimmed mean. $\bar{x}$ is not a robust estimator as it is sensitive to outliers. Nevertheless $\bar{x}$ is the preferred estimator of μ. $\bar{x}$ is the estimator with the smallest variance.

What is the variance of $\bar{x}$?

We had $\text{Var}[x_i] = \sigma^2$.

$$\begin{aligned}
\text{Var}[\bar{x}] &= \text{Var}\left[\frac{x_1 + \cdots + x_n}{n}\right] \\
&= \frac{1}{n^2}\left(\text{Var}[x_1 + \cdots + x_n]\right) \\
&= \frac{1}{n^2}\left(\text{Var}[x_1] + \cdots + \text{Var}[x_n]\right) \qquad \text{Independence!} \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned} \tag{2.6}$$

x_i or $\bar{x}$?

$$\mathcal{E}[x_i] = \mu \qquad \text{Var}[x_i] = \sigma^2 \tag{2.7}$$

$$\mathcal{E}[\bar{x}] = \mu \qquad \text{Var}[\bar{x}] = \frac{\sigma^2}{n} \tag{2.8}$$

$$\text{If } x \sim \mathcal{N}(0,1), \text{ then } \bar{x} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$$

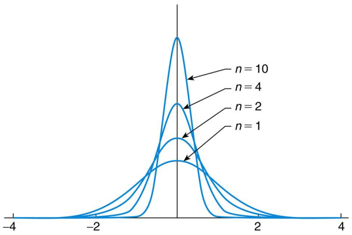


Fig. 2.1: The distribution of $\bar{x}$

The sample mean depends on n. Therefore we must sample more and gain accuracy!

Intuitively, we desire replication of results.

The population size can matter

For an infinite population, $\text{Var}[\bar{x}] = \sigma^2/n$. For a finite population of size N, we need a correction factor.

More on this in a separate handout.

$$\text{Var}[\bar{x}] = \frac{\sigma^2}{n}\frac{N-n}{N-1} \tag{2.9}$$

The Central Limit Theorem

Theorem 2.1. *The sum of a large number of independent random variables has a distribution that is approximately normal.*

Let $x_1, \ldots, x_n$ be i.i.d. with mean μ and variance σ^2. For large n

$$(x_1 + \cdots + x_n) \sim \mathcal{N}(n\mu, n\sigma^2) \tag{2.10}$$

$$P\left(\frac{(x_1 + \cdots + x_n) - n\mu}{\sigma\sqrt{n}}\right) \approx P(z < x) \tag{2.11}$$

Consider the total obtained on rolling several (n) dice (Fig. 2.2). As n increases, the distribution changes to a Gaussian curve!

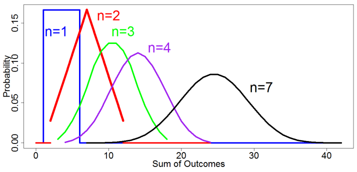


Fig. 2.2: Rolling n dice.

CLT and the Binomial

Let $X = x_1 + \cdots + x_n$, where x_i is a Bernoulli variable.

$$\mathcal{E}\left[x_i\right] = p \qquad \text{Var}\left[x_i\right] = p(1-p) \tag{2.12}$$

$$\mathcal{E}\left[X\right] = np \qquad \text{Var}\left[X\right] = np(1-p) \tag{2.13}$$

CLT suggests that for large n

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1) \tag{2.14}$$

The Binomial can be approximated with a normal if $np(1-p) \geq 10$. For $p = 0.5$, $npq \geq 10$ implies $n \geq 40$! In general, the normal approximation may be applied if $n \geq 30$. So keep sampling!

Remember: even if an individual measurement does not follow a Normal distribution, the measured sample mean does.

Sample variance

The population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{2.15}$$

The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.16}$$

Why $n-1$?

We had $\mathcal{E}\left[\bar{x}\right] = \mu$. But is $\mathcal{E}\left[S^2\right] = \sigma^2$? We compute $\mathcal{E}\left[S^2\right]$ as follows:

$$\begin{aligned}(x_i - \bar{x})^2 &= (x_i - \mu + \mu - \bar{x})^2 \\ &= (x_i - \mu)^2 + (\mu - \bar{x})^2 + 2(x_i - \mu)(\mu - \bar{x}) \\ \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \mu)(\mu - \bar{x}) \end{aligned} \tag{2.17}$$

The middle term is ...

$$\sum_{i=1}^{n}(\mu - \bar{x})^2 = n(\mu - \bar{x})^2$$

The last term is ...

$$2\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x}) = -2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu) = -2n(\bar{x}-\mu)^2$$

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i-\mu)^2 + n(\mu-\bar{x})^2 - 2n(\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\\ \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} &= \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} - (\bar{x}-\mu)^2\end{aligned}$$

Taking the expectation on each side

$$\mathcal{E}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right] = \mathcal{E}\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right] - \mathcal{E}\left[(\bar{x}-\mu)^2\right] \tag{2.18}$$

But,

$$\mathcal{E}\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right] = \sigma^2 \tag{2.19}$$

$$\mathcal{E}\left[(\bar{x}-\mu)^2\right] = \mathrm{Var}\left[\bar{x}\right] = \frac{\sigma^2}{n} \tag{2.20}$$

This implies

$$\mathcal{E}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right] = \sigma^2 - \frac{\sigma^2}{n} = \frac{(n-1)\sigma^2}{n} \tag{2.21}$$

We have a biased estimator! This means that

$$\mathcal{E}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right] = \sigma^2 \tag{2.22}$$

Therefore we define an unbiased estimator s^2 as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \tag{2.23}$$

We lost a degree of freedom. Where?

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{n} \qquad\qquad s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

2.2 Point and Interval Estmation

Point estimates

Given $x_i \sim \mathcal{N}(\mu,\sigma^2)$, the preferred point estimate of μ is $\bar{x}$. The preferred point estimate of σ^2 is s^2. We know how $\bar{x}$ behaves:

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad\qquad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{2.24}$$

We do not know how s^2 behaves yet. Similarly, intuition leads us to the best estimate of the binomial proportion p as $\hat{p} = X/n$. The maximum likelihood estimation approach below provides a method for the determination of parameter point estimates.

Optimal estimation approaches

The MLE of a parameter θ is that value of θ that maximizes the likelihood of seeing the observed values, i.e. $P(D|\theta)$. Note that instead we could desire to maximize $P(\theta|D)$: i.e. we choose the most probable parameter value given the data that we have already seen. This is Maximum a posteriori (MAP) estimation. The MAP estimate is typically more difficult to obtain because we need to rewrite it in terms of $P(D|\theta)$ using Bayes law, and $p(\theta)$ needs to be specified. Also note that if $P(\theta)$ is effectively a constant (i.e. θ is a uniform RV), then MLE = MAP. There are other approaches towards obtaining estimates of parameters including the confusingly named Bayesian estimation (different from MAP), maximum entropy etc. In general note that we want estimates with small variances, and hence minimum variance estimates are popular. In the discussion that follows, we use almost always use MLE; we will however encounter MAP at least once later.

MLE of the Binomial parameter p:

Let X = random variable describing a coin flip, $X = 1$ if heads, $X = 0$ if tails. For one coin flip,

As before, the Bernoulli and the Binomial.

$$f(x) = p^x(1-p)^{1-x} \tag{2.25}$$

For n coin flips,

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{2.26}$$

The log-likelihood may be written as

$$l(p|x_i, ..., x_n) = \sum_{i=1}^{n} [x_i \ln p + (1-x_i)\ln(1-p)] \tag{2.27}$$

The MLE of $p = \hat{p}$ and is obtained from

$$\frac{d\,l(p)}{dp} = 0 \qquad \Rightarrow \qquad \frac{\sum x_i}{\hat{p}} - \frac{\sum(1-x_i)}{1-\hat{p}} = 0$$

which can be rearranged to give

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.28}$$

The answer here is obvious. MLE does this quite often: it gives us intuitive answers. However as the next example shows, MLE could be biased.

The MLE of the binomial proportion (X heads in n tosses) is

$$\hat{p} = \frac{X}{n} \qquad \text{and} \qquad \mathcal{E}[\hat{p}] = p \tag{2.29}$$

An MLE perspective on $\bar{x}$ and s^2, for a Gaussian

Let $D = \{x_1, \ldots, x_n\}$ be a dataset of i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$. Then the multivariate density function is

$$f(D) = \prod_{i=1}^{n} \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \tag{2.30}$$

$f(D)$ is the likelihood of seeing the data. $l = \ln[f(D)]$ is the log likelihood of seeing the data. Which estimates of μ and σ^2 would maximize the likelihood of seeing the data? Set

$$\frac{\partial l}{\partial \mu} = 0 \qquad \frac{\partial l}{\partial \sigma^2} = 0 \tag{2.31}$$

and solve for $\hat{\mu}$ and $\hat{\sigma^2}$.

$$\frac{\partial l}{\partial \mu} = 0 \qquad \Rightarrow \qquad \frac{\partial \sum_{i=1}^{n}(x_i-\mu)^2}{\partial \mu} = 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_i x_i}{n} = \bar{x} \tag{2.32}$$

The MLE of μ is $\hat{\mu} = \bar{x}$. The MLE of σ^2 however is a surprise:

$$\frac{\partial l}{\partial \sigma^2} = 0 = -\frac{n}{2\sigma^2} + \frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial \sigma^2}$$

$$\begin{aligned}
\frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial \sigma^2} &= \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \\
&= \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}\left[(x_i-\bar{x})+(\bar{x}-\mu)\right]^2\right] \\
&= \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})\right] \\
&= \frac{1}{2\sigma^4}\left[nS^2 + n(\bar{x}-\mu)^2\right] \qquad \left[\text{where } S^2 \text{ is defined as } = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma^2} = 0 &= -\frac{n}{2\sigma^2} + \frac{\partial\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{\partial \sigma^2} \\
0 &= -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left[S^2 + (\bar{x}-\mu)^2\right]
\end{aligned}$$

The MLE of σ^2 is biased!

But $\bar{x} = \mu$ from before, and hence

$$\sigma^2_{\text{ML}} = S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \tag{2.33}$$

Interval Estimation

We have point estimates of μ and σ^2. But remember that $\bar{x}$ and σ^2 are random variables. We could use bounds: for example, $\bar{x} = \mu \pm \delta$. δ should depend on σ^2.

Chebyshev's inequality for a random variable x
Theorem 2.2.

$$P\left(|x_i - \mu| > k\sigma\right) \le \frac{1}{k^2} \tag{2.34}$$

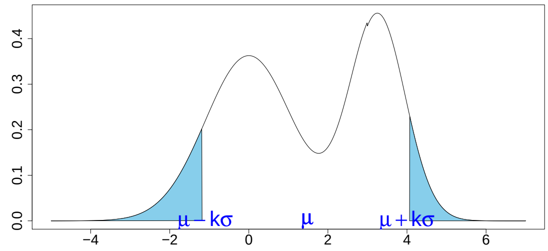


Fig. 2.3: Chebyshev's theorem

$$\begin{aligned}
\sigma^2 = \mathcal{E}\left[(x-\mu)^2\right] &= \sum_{R_1,R_2,R_3}\left(x_i - \mathcal{E}[x]\right)^2 p(x_i) \\
&\ge \sum_{R_1}\left(x_i - \mathcal{E}[x]\right)^2 p(x_i) + \sum_{R_3}\left(x_i - \mathcal{E}[x]\right)^2 p(x_i) \\
&\ge \sum_{R_1} k^2\sigma^2 p(x_i) + \sum_{R_3} k^2\sigma^2 p(x_i) \\
\frac{1}{k^2} &\ge \sum_{R_1} p(x_i) + \sum_{R_3} p(x_i)
\end{aligned}$$

$$P(|x_i - \mu| > k\sigma) \le \frac{1}{k^2}$$

This applies to ANY distribution. For $k = 2$,

- Chebyshev's theorem suggests: $P(|x_i - \mu| > 2\sigma) \le 0.25$.
- For a Gaussian, $P(|x_i - \mu| > 2\sigma) \le 0.05$.

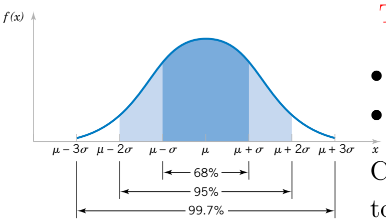


Fig. 2.4: Intervals on a Gaussian

Clearly, if the density function is known, we should be using it instead of Chebyshev's theorem, to determine probabilities.

Chebyshev's inequality for the derived variable $\bar{x}$

$$P\left(|\bar{x}-\mu| > k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2} \tag{2.35}$$

Let $k\sigma/\sqrt{n} = \varepsilon =$ some tolerance.

$$P(|\bar{x}-\mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \qquad \Rightarrow \qquad P(|\bar{x}-\mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \tag{2.36}$$

This is the law of large numbers: For a given tolerance, as $n\uparrow$, then $\bar{X} \to \mu$.

Which estimate intervals do we want?

Chebyshev's inequality gives us a loose bound. We want better estimates. For a normal distribution,

1. interval estimate of μ with σ^2 known,
2. interval estimate of μ with σ^2 unknown,
3. interval estimate of σ^2 with μ unknown

For a binomial distribution,

4. interval estimate of p

1. Normal: interval for μ, σ^2 known

We have n normally distributed points.

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad\qquad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

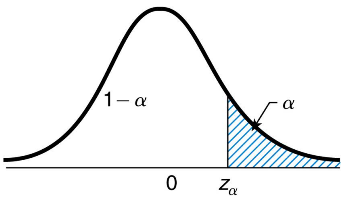


Fig. 2.5: One sided interval

Recall the z_α notation for a threshold:

$$P\left(z > Z_\alpha\right) = \alpha$$

For a two-sided interval,

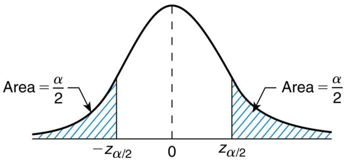


Fig. 2.6: Two sided intervals

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha \tag{2.37}$$

$$P\left(-1.96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \tag{2.38}$$

We can rearrange the inequalities in

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

to get

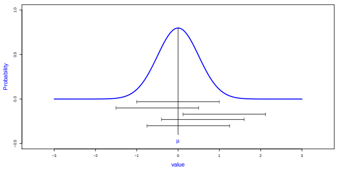


Fig. 2.7: Confidence intervals

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \tag{2.39}$$

So what does a confidence interval $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ mean?: The maximum error of the estimate is $|\bar{X}-\mu| < z_{\alpha/2}\sigma/\sqrt{n}$ with probability $1-\alpha$. What does a confidence interval mean for a Frequentist? The two-sided confidence interval at level $100(1-\alpha)$ was

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.40}$$

We can have lower and upper one sided intervals.

$$\left[\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\right] \tag{2.41}$$

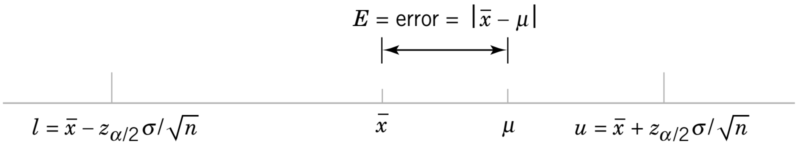
$$\left[-\infty,\, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right] \tag{2.42}$$

What should the sample size (n) be, if we desire $\bar{x}$ to approach μ to within a desired level of confidence (i.e. given α)? Rearrange and solve for n:

$$n = \left(\frac{z_{\alpha/2}\sigma}{|\bar{x}-\mu|}\right)^2 \tag{2.43}$$

Does the dependency of n on the various terms make sense? You need more samples if

1. you want $\bar{x}$ to come very close to μ,


E = error = $|\bar{x} - \mu|$
$l = \bar{x} - z_{\alpha/2}\sigma/\sqrt{n}$
$\bar{x}$
μ
$u = \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$

Intervals on the Gaussian

2. σ is large,
3. $z_{\alpha/2}$ is increased (i.e. α is decreased)

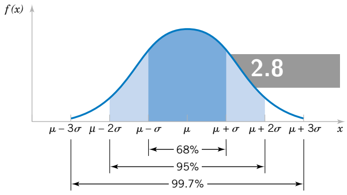


Fig. 2.9: Effect of α

2. Normal: interval for μ, σ^2 unknown

The two sided interval for μ, when σ^2 is known is $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. What happens when σ^2 is unavailable? We must estimate σ^2. It's point estimate is s^2. However, there's a problem w.r.t. interval estimation:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \text{ follows the normal } \mathcal{N}(0,1)$$

$$\text{But } \frac{\bar{x}-\mu}{s/\sqrt{n}} \text{ does not follow } \mathcal{N}(0,1)$$

In fact, we need a new sampling distribution to describe how $(\bar{x}-\mu)/(s/\sqrt{n})$ varies. We need to acknowledge that our estimate s^2 is a random variable (RV), with its own distribution centered around mean σ^2. Therefore we have a range of values for s^2 which could be used to substitute for σ^2 in $\mathcal{N}(\mu,\sigma^2)$ and therefore we must average across several Gaussians, which have the same mean but with different variances. The new distribution must have the following properties:

- The Gaussians are symmetric about μ: therefore a weighted average of Gaussians must be symmetric about μ.
- Tail: it should have a thick tail, for small n: greater variability than $\mathcal{N}$.
- As $n \to \infty$, we should be back to $\mathcal{N}(\mu,\sigma^2)$.
- The distribution is a function of sample size n. We have a family of curves!

William Gosset: a 'chemist' at the Guinness breweries in Dublin, came up with the Student's t-distribution!

We know of $\mathcal{N}(x|\mu,\sigma^2)$. But σ^2 is unknown, and if we estimate it, we end up with a distribution of variance values. It is easier to work with the precision τ:

$$\tau = \frac{1}{\sigma^2} \tag{2.44}$$

The univariate Normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{\tau^{1/2}}{\sqrt{2\pi}}\exp\left[-\frac{\tau}{2}(x-\mu)^2\right] \tag{2.45}$$

We have a dependency on τ of the form

$$\frac{\tau^{1/2}}{\sqrt{2\pi}}\exp\left[-\frac{\tau}{2}(x-\mu)^2\right]$$

The Gamma distribution is

$$Gam(x|\alpha,\lambda)=\frac{\lambda^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}\exp\left(-\lambda x\right) \qquad x>0, \qquad \alpha>0, \qquad \lambda>0 \tag{2.46}$$

where

$$\Gamma(\alpha)=\int_0^{\infty}e^{-y}y^{\alpha-1}dy \tag{2.47}$$

$\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$. If α is an integer, $\Gamma(n)=(n-1)!$. For a Gamma variable, the moment generating function is

$\alpha = 1$: $\Gamma(1) = 1$. The Gamma distribution becomes an exponential distribution.

$$\Phi(t)=\mathcal{E}\left[e^{tx}\right]=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty}e^{tx}e^{-\lambda x}x^{\alpha-1}dx=\left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \tag{2.48}$$

$$\mathcal{E}\left[x\right]=\Phi^{'}(0)=\frac{\alpha}{\lambda} \tag{2.49}$$

$$\text{Var}\left[x\right]==\Phi^{''}(0)-(\mathcal{E}\left[x\right])^2=\frac{\alpha}{\lambda^2} \tag{2.50}$$

The precision τ follows a Gamma distribution. We had

$$\mathcal{N}(x|\mu,\sigma^2)=\mathcal{N}(\mu,\tau^{-1}) \tag{2.51}$$

We want the average of several distributions:

$$\mathcal{E}_{\tau}\left[\mathcal{N}(\mu,\tau^{-1})\right] \tag{2.52}$$

which is

$$\int_0^{\infty}\mathcal{N}\left(x|\mu,\tau^{-1}\right)Gam(\tau|\alpha,\lambda)d\tau \tag{2.53}$$

This averaged distribution, can be shown to be (details in another handout on Moodle)

$$\frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)}\left(\frac{1}{2\pi\lambda}\right)^{1/2}\left[1+\frac{(x-\mu)^2}{2\lambda}\right]^{-\alpha-\frac{1}{2}} \tag{2.54}$$

This is the Student's t-distribution, $St(x|\mu,\alpha,\lambda)$ where 2α signifies the degrees of freedom. Notice the symmetry about 0 in Figs 2.11 and 2.12. Quantiles corresponding to the 95% confidence interval for the t-distribution, as a function of the degrees of freedom, are shown in Table 2.1. For $n\geq 30$, $t_{\nu}\approx\mathcal{N}(0,1)$. Now

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\text{ follows the }\mathcal{N}(0,1)\text{ curve, but }\frac{\bar{x}-\mu}{s/\sqrt{n}}\text{ follows the }t_{n-1}\text{ curve}$$

The interval for μ when σ^2 was known, was

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

The interval for μ when σ^2 is unknown is

$$\bar{x}\pm\frac{s}{\sqrt{n}}t_{\alpha/2,n-1} \tag{2.55}$$

Similarly, lower and upper one sided intervals are

$$\left[\bar{x}-t_{\alpha,n-1}\frac{s}{\sqrt{n}},\ \infty\right] \tag{2.56}$$

$$\left[-\infty,\ \bar{x}+t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right] \tag{2.57}$$

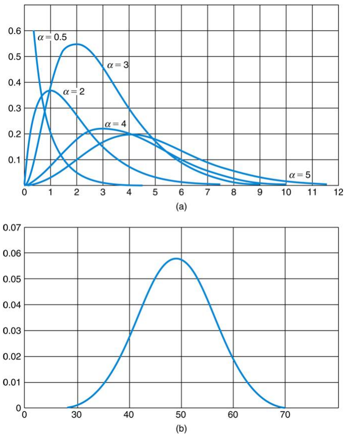


Fig. 2.10: Gamma distribution

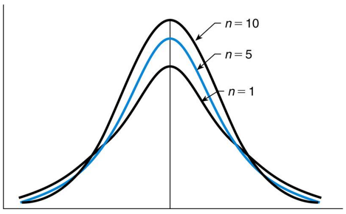


Fig. 2.11: The t distribution

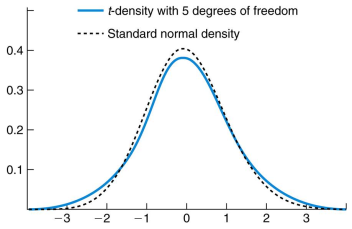


Fig. 2.12: The t and $\mathcal{N}(0,1)$ distributions

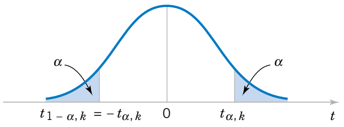


Fig. 2.13: Intervals on the t distribution

ν	$t_{0.025,\nu}$
4	2.776
9	2.262

Example 2.3. A fuse manufacturer states that with an overload, fuses blow in 12.40 minutes on average. As a test, take 20 sample fuses and subject them to an overload. A mean time of 10.63 and a standard deviation of 2.48 minutes is obtained. Is the manufacturer right?

Soln:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{10.63 - 12.40}{2.48/\sqrt{20}} = -3.19$$

From a t table,

$$P(t_{19} > 2.861) = 0.005 \qquad \Rightarrow \qquad P(t < -2.861) = 0.005$$

Hence the manufacturer is wrong and the mean time is probably (with at least 99% confidence) below 12.40 minutes.

3. Normal: interval for σ^2

If $(x_1, \dots, x_n) \sim \mathcal{N}(\mu, \sigma^2)$. The point estimate of σ^2 is s^2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \text{and } \mathcal{E}\left[s^2\right] = \sigma^2 \tag{2.58}$$

We have

$$z_i = \frac{(x_i - \mu)}{\sigma} \sim \mathcal{N}(0,1) \tag{2.59}$$

Let

$$X = z_1^2 + \cdots + z_n^2 \tag{2.60}$$

X follows a χ^2 distribution with n degrees of freedom.

$$X \sim \chi_n^2 \tag{2.61}$$

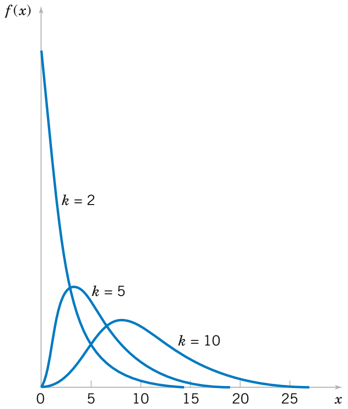


Fig. 2.14: The χ^2 distribution. It is not symmetric. The skew ↓ as n ↑.

The χ^2 with n degrees of freedom is related to the Gamma distribution. The χ^2 distribution has x always > 0 and is skewed to the right. We have

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} \sim \chi_n^2 \tag{2.62}$$

When we substitute μ with $\bar{x}$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} \sim \chi_{n-1}^2 \tag{2.63}$$

But

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2 \qquad \Rightarrow \qquad s^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2 \tag{2.64}$$

We compute intervals, using quantiles, as before. No symmetry ⇒ we need to evaluate both quantiles for a two sided interval.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2 \tag{2.65}$$

$$P\left(\frac{(n-1)s^2}{\chi_{\alpha/2,n-1}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2,n-1}^2}\right) = 1 - \alpha \tag{2.66}$$

The 100% × $(1 - \alpha)$ confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi_{\alpha/2,n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2,n-1}^2}\right] \tag{2.67}$$

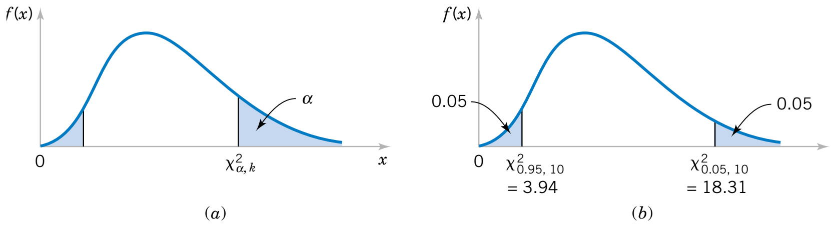


Fig. 2.15 Intervals on the χ^2 distribution

One sided intervals can be obtained similarly:

$$\left[0, \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}\right] \tag{2.68}$$

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}}, \infty\right] \tag{2.69}$$

Notice how we standardize:

- We standardize x_i as

$$\frac{x_i-\mu}{\sigma}$$

- We standardize $\bar{x}$ as

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

- We standardize s^2 as

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

4. Binomial: interval for p

Assume n trials, X positive outcomes. Then,

$$X \sim Bi(n,p) \qquad \mathcal{E}\left[X\right] = np \qquad \mathrm{Var}\left[X\right] = np(1-p) \tag{2.70}$$

As n increases, using CLT,

$$X \sim \mathcal{N}(np, np(1-p)) \tag{2.71}$$

Therefore

$$\frac{X-np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1) \tag{2.72}$$

To obtain the interval, observe that

$$X \sim \mathcal{N}(np, np(1-p)) = \mathcal{N}\left(\mu, \sigma^2\right) \tag{2.73}$$

The Gaussian interval:

$$P\left(-z_{\alpha/2} < \frac{x-\mu}{\sigma} < z_{\alpha/2}\right) = 1-\alpha$$

The confidence interval therefore should be

$$P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right) \approx 1-\alpha \tag{2.74}$$

We need to replace p with $\hat{p} = X/n = p_{\mathrm{ML}}$

$$P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right) \approx 1-\alpha \tag{2.75}$$

Multiply throughout by $\sqrt{n\hat{p}(1-\hat{p})}$. Also, since $X = n\hat{p}$,

$$P\left(-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < n\hat{p} - np < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})}\right) \approx 1-\alpha \tag{2.76}$$

Rearrange.

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha \tag{2.77}$$

The interval is

$$\left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{2.78}$$

Let b be the width of a $100\% \times (1-\alpha)$ confidence interval:

$$b = 2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.79}$$

Solving for n gives

$$n = \hat{p}(1-\hat{p})\left[\frac{2z_{\alpha/2}}{b}\right]^2 \tag{2.80}$$

We can get an upper bound on n for $\hat{p} = 1/2$:

$$n = \frac{1}{4}\left(\frac{2z_{\alpha/2}}{b}\right)^2 \tag{2.81}$$

$$n = \frac{1}{4}\left(\frac{2z_{\alpha/2}}{b}\right)^2$$

This can be tedious.

Remember that the normal approximation to the binomial holds for $n\hat{p}(1-\hat{p}) \geq 10$. If $n\hat{p}(1-\hat{p}) < 10$, then we cannot use the normal approximation. The exact (but hard) way to find the interval $[p_1, p_2]$ would be

$$\begin{aligned} P(X \geq x | p = p_1) &= \frac{\alpha}{2} = \sum_{k=x}^{n} {}^{n}C_k p_1^k (1-p_1)^{n-k} \\ P(X \leq x | p = p_2) &= \frac{\alpha}{2} = \sum_{k=0}^{x} {}^{n}C_k p_2^k (1-p_2)^{n-k} \end{aligned} \tag{2.82}$$

2.3 Hypothesis Testing

The hypothesis can be about

- the value of a parameter (or parameters).
- a qualitative statement.

The null hypothesis is about population parameters (and not about samples).

$$H_0 : \mu = \mu_0 \qquad H_0 : \sigma^2 = \sigma_0^2 \tag{2.83}$$

To test a hypothesis about a population parameter, we need

- a test statistic computed from sampled data, and
- the corresponding distribution on which to determine an interval estimate.

How do we formulate the test hypothesis?

The null hypothesis, H_0, is the specific hypothesis we wish to test.

$$H_0 : X \text{ is guilty} \tag{2.84}$$

Every hypothesis can have one or more alternates.

$$H_0 : X \text{ is guilty} \qquad H_1 : X \text{ is innocent}$$

The hypothesis $H_0 : \theta = \theta_0$ has many alternatives:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_1 &: \theta < \theta_0 \\ H_1 &: \theta > \theta_0 \\ H_1 &: \theta = \theta_1 \, (\neq \theta_0) \end{aligned}$$

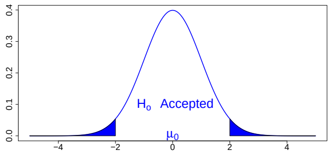


Fig. 2.16: Hypothesis acceptance

Do we need to pay attention to H_1? Yes. H_1 influences the interval estimate. To see that we need to understand what α means. But before that, why should H_0 be the equality statement? Because the moment we say something like $H_0 : \mu = \mu_0$ we have a specific distribution to work with! The equality statement carries the burden of proof. The legal system prefers that we try and disprove $H_0 : X$ is innocent. Typically, the accused is innocent (H_0) until proven guilty (H_1) beyond reasonable doubt (α).

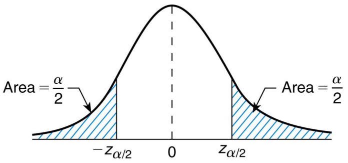


Fig. 2.17: The error α

What is α?

The error of the estimate is $|\bar{x} - \mu|$.
This maximum error is $< z_{\alpha/2}\sigma/\sqrt{n}$ with probability $1 - \alpha$ There are 4 possible outcomes to a test of H_0 vs. H_1.

1 Accept H_0 as true, when in fact it is true.
2 Accept H_0 as true, when in fact it is not true (i.e. H_1 is true).
3 Reject H_0 as true, when in fact H_0 is true.
4 Reject H_0 as true, when in fact H_0 is not true (i.e. H_1 is true).

	Truth	
	H_0	H_1
Decision	H_0 true, H_0 accepted	H_1 true, H_0 accepted
Taken	H_0 true, H_1 accepted	H_1 true, H_1 accepted

Type I error: H_0 true, but H_1 accepted $= \alpha$

Remember that α = significance level of the test applied.

Where there's an α...

...there's a β. Type II error: H_1 true, but H_0 accepted $= \beta$. Assuming $H_1 : \mu = \mu_1$, with $\mu_1 > \mu_0$, $1-\beta$ is the power of the test. β depends on μ_1. Increase $|\mu_0 - \mu_1|$ and $1-\beta$ increases $\Rightarrow$ "powerful" test. We accept H_0 in the range

$$= \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2} = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.85}$$

But β is the area in this range, under H_1. We need to convert the thresholds to standardized coordinates first, this time under H_1. Under H_1,

$$\bar{x} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right) \tag{2.86}$$

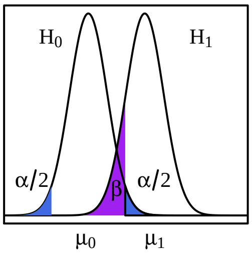


Fig. 2.18: α and β

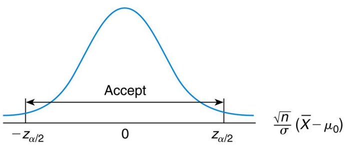


Fig. 2.19: The acceptance interval

The location $\mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$ becomes (under H_1)

$$= \frac{(\mu_0 - z_{\alpha/2}\sigma/\sqrt{n}) - \mu_1}{\sigma/\sqrt{n}} = \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2} \tag{2.87}$$

Similarly, the other threshold is

$$= \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2} \tag{2.88}$$

Therefore

$$\begin{aligned}\beta &= P\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2} \le z \le \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right)\\ &= \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2}\right)\end{aligned} \tag{2.89}$$

If $\mu_1 > \mu_0$, then the second term ≈ 0.

$$\beta \approx \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) = P(z > z_\beta) = P(z < -z_\beta) = \Phi(-z_\beta) \tag{2.90}$$

$$-z_\beta \approx (\mu_0 - \mu_1)\frac{\sqrt{n}}{\sigma} + z_{\alpha/2} \tag{2.91}$$

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{(\mu_1 - \mu_0)^2} \tag{2.92}$$

The sample size depends on

1. $|\mu_0 - \mu_1|$. As $|\mu_0 - \mu_1|$ increases, n decreases.

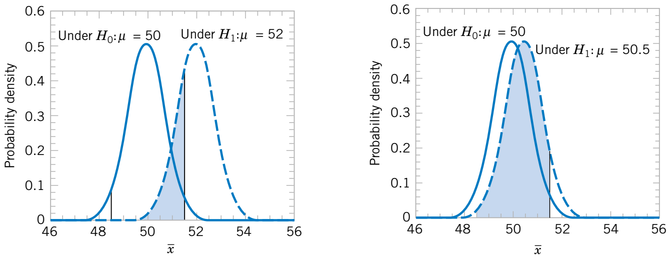


Fig. 2.20 Effect of $|\mu_0 - \mu_1|$

2. σ^2. As σ^2 increases, n increases.

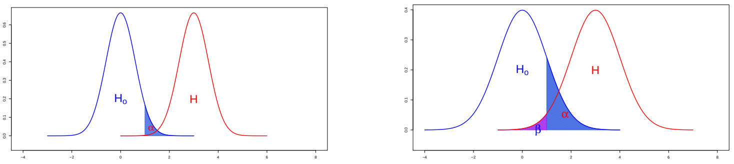


Fig. 2.21 Effect of σ^2

3. α. As α decreases, $z_{\alpha/2}$ increases, hence n increases.

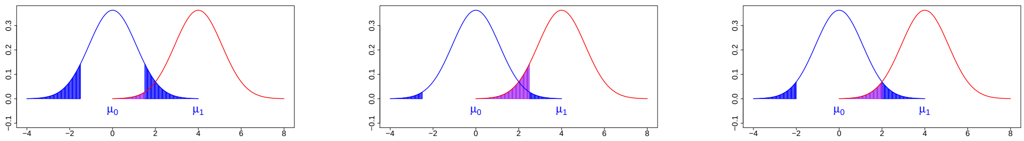


Fig. 2.22 Effect of α

4. β. As required power increases ($1 - \beta$ increases), n increases.

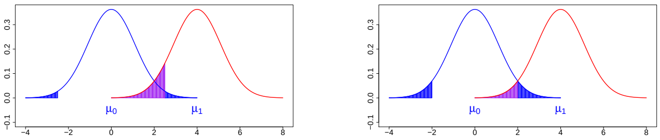
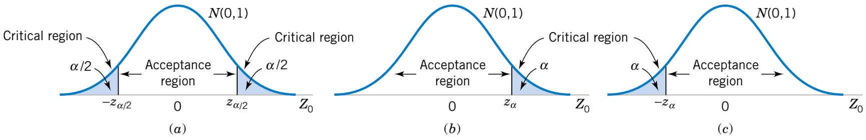


Fig. 2.23 Effect of β

One-sided vs. two-sided tests

Consider the null hypothesis $H_0 : \mu = \mu_0$: When $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$, we should



$H_1 : \mu \neq \mu_0$ $H_1 : \mu > \mu_0$ $H_1 : \mu < \mu_0$

Fig. 2.24 One-sided vs. two-sided tests

Which is more conservative: a two-sided test or a one-sided test?

$$\text{accept} \quad H_0 \quad \text{if} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \tag{2.93}$$

$$\text{reject} \quad H_0 \quad \text{if} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \tag{2.94}$$

p-value

The p-value is the significance level when no decision can be made between accepting and rejecting H_0. Find the probability of exceeding the test statistic (in absolute value) on a unit normal.

$$p\text{-value } = p\left(|z| > \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right) \tag{2.95}$$

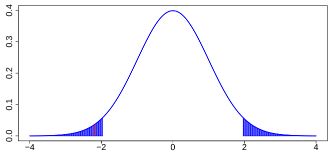


Fig. 2.25: The p value

p-value and one-sided tests

- For $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$, we would reject if the test statistic, t is $> z_{\alpha/2}$.

$$p\text{-value } = 2 \times p(z \geq |t|)$$

- For $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$, we would reject if the test statistic, t is $> z_\alpha$.

$$p\text{-value } = p(z \geq t)$$

- For $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$, we would reject if the test statistic, t is $< -z_\alpha$.

$$p\text{-value } = p(z \leq t)$$

Why is it important to report the p-value? Why not just report the test result (yes/no) based on the chosen α?

When you perform a hypothesis test, you should report

1. the significance level of the test that you chose,
2. AND the p-value associated with your test statistic.

A problem solving tip

For $\mu_1 > \mu_0$, and for a two-sided test,

$$\beta \approx \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) \tag{2.96}$$

Observe that

$$\beta = f(|\mu_0 - \mu_1|,\, \sigma^2,\, n,\, \alpha) \tag{2.97}$$

The sample size n is

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{(\mu_1 - \mu_0)^2} \tag{2.98}$$

and depends on

$$n = f(|\mu_0 - \mu_1|,\, \sigma^2,\, \alpha,\, \beta) \tag{2.99}$$

Most problems require us to solve for one of the following in terms of the rest:

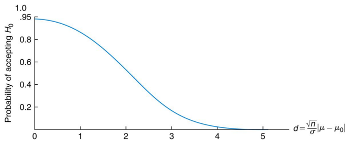
$$|\mu_0 - \mu_1|,\, \sigma^2,\, \alpha,\, \beta,\, n$$

β, power and rejection of H_0

As we increase μ_1, β decreases. The curve described by $\beta(\mu_1)$ is called an operating characteristic, (OC). For a fixed value of α, plot

$$\beta(\mu_1) \quad \text{vs.} \quad \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} \tag{2.100}$$

The Operating Characteristic curve for a two-sided test with $\alpha = 0.05$.



$$\beta(\mu_1) \quad \text{vs.} \quad \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}$$

Fig. 2.26: The OC curve.

One sided tests and β

We had, when $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$,

$$\beta(\mu_1) = P_\mu\,(\text{accepting } H_0 \quad | \quad H_1 : \mu = \mu_1 \text{ is correct}) \tag{2.101}$$

We switch to the more general notation, replacing μ_1 with μ

$$beta(\mu) = P_\mu\,(\text{accepting } H_0 \quad | \quad H_1 \text{ is correct}) \tag{2.102}$$

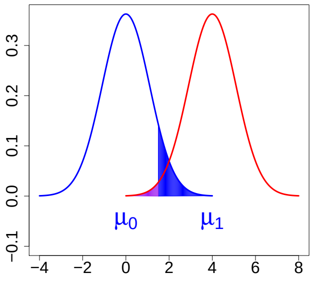


Fig. 2.27: Finding β.

$$\begin{aligned}
\beta(\mu) &= P_\mu\left(\bar{x} \le \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&= P\left(z \le \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right), \qquad\qquad z \sim \mathcal{N}(0,1) \\
&= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \qquad (2.103)
\end{aligned}$$

Compare this against β for a two sided interval: $z_{\alpha/2}$ has been replaced with α, and μ_1 has been replaced with the more generic μ. As its argument increases, Φ increases. Therefore as μ increases, $\beta(\mu)$ decreases. The larger μ is, the less likely it would be to conclude that

- $\mu = \mu_0$,
- or for that matter $\mu \le \mu_0$.

$\beta(\mu_0) = 1 - \alpha$

H_0 and H_1

To test $H_0 : \mu{=}\mu_0$ and $H_1 : \mu > \mu_0$, we compared

$$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \qquad \text{vs.} \qquad z_\alpha$$

If instead, we wished to test $H_0 : \mu{\le}\mu_0$ and $H_1 : \mu > \mu_0$, what changes must we make to the test procedure? How did $H_0 : \mu = \mu_0$ influence the testing procedure?

- We put down the H_0 curve centered around μ_0, and then
- chose α and identified an interval based on it.

We meant that "the probability of rejection of H_0, should be (the small valued) α". As we relax H_0 to now be $\mu \le \mu_0$ we need to ensure that the probability of rejection of H_0 is never greater than α. But the probability of accepting H_0 for a given μ is $\beta(\mu)$. Therefore, for the probability of rejection to be never greater than α, we would need

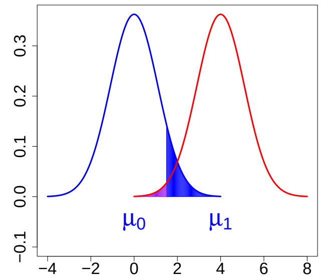


Fig. 2.28: β and α

$$1-\beta(\mu) \le \alpha \qquad \text{for all} \qquad \mu \le \mu_0$$

which implies that we would need

$$\beta(\mu) \ge 1-\alpha \qquad \text{for all} \qquad \mu \le \mu_0$$

But, as seen before, $\beta(\mu)$ decreases as μ increases. Therefore $\beta(\mu) \ge \beta(\mu_0)$. But $\beta(\mu_0) = 1-\alpha$, and therefore we confirm that

$$\beta(\mu) \ge= 1-\alpha \quad \text{for all} \quad \mu \le \mu_0$$

Therefore. a test for $H_0 : \mu = \mu_0$ at level α also works for $H_0 : \mu \le \mu_0$ at level α.

2.4 Parametric hypothesis testing

Overview of Parametric Tests discussed

1. Testing the mean of a Normal Population variance known ('z-Test')
2. Testing the mean of a Normal Population variance unknown ('t-Test')
3. Testing the variance of a Normal Population ('χ^2-Test')
4. Testing the Binomial proportion
5. Testing the Equality of Means of Two Normal Populations, variances known
6. Testing the Equality of Means of Two Normal Populations, variances unknown but assumed equal
7. Testing the Equality of Means of Two Normal Populations, variances unknown and found to be unequal
8. Testing the Equality of Variances of Two Normal Populations ('F-Test')
9. Testing the Equality of Means of Two Normal Populations, the Paired t-Test.

The approach for any Parametric Test

Step 1. State the hypothesis.

a. Identify the random variable and the relevant parameter to be tested.
b. Clearly state H_0 and H_1.
c. Identify the distribution that applies under H_0.

Step 2. Identify the significance level (α) at this stage.

Step 3. Based on H_1 and α identify the critical threshold(s).

a. The relevant (standardized) sampling distribution needs to be identified knowing sample size/degrees of freedom.

b. One-sided or two-sided test?

Step 4. Compute the test statistic from the sampled data.

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

1. The z-Test: One sample test for μ when the variance is known

Step 1. State the hypothesis.

$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$

Under H_0, $\bar{x} \sim \mathcal{N}(\mu_0, \sigma^2)$

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

At $\alpha = 0.05$, on the Standard Normal, for a two-sided test, $z_{\alpha/2} = 1.96$.

Step 4. Compute the test statistic from the sampled data.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad \text{vs.} \qquad \mathcal{N}(0,1)$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

2. The t-Test: One sample test for μ when the variance is unknown

Step 1. State the hypothesis.

$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$

Under H_0, $\bar{x} \sim \mathcal{N}(\mu_0, \sigma^2)$. σ^2 is unknown. Standardize and use $t_{\alpha/2,n-1}$ curve.

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

Use $\alpha = 0.05$, on the t_{n-1} curve

Step 4. Compute the test statistic from the sampled data.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{vs.} \qquad t_{n-1}$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

Example 2.4. 12 rats are subjected to forced exercise. Each weight change (in g) is the weight after exercise minus the weight before. Does exercise cause weight loss?

1.7, 0.7, -0.4, -1.8, 0.2, 0.9,
-1.2, -0.9, -1.8, -1.4, -1.8, -2.0

Soln:

Step 1. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.
Step 2. Use $\alpha = 0.05$
Step 3. $t_{0.025,11} = 2.201$
Step 4. $n = 12$. $\bar{x} = -0.65$ g and $s^2 = 1.5682$ g^2.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = -1.81$$

Step 5. Since $|t| < |t_{\alpha/2,n-1}|$ then H_0 cannot be rejected.
Step 6. p-value = ?

Example 2.5. The level of metabolite X measured in 12 patients 24 hours after they received a newly proposed antibiotic was 1.2 mg/dL. If the mean and standard deviation of X in the general population are 1.0 and 0.4 mg/dL, respectively, then, using a significance level of 0.05, test if the level of X in this group is different from that of the general population. What happens if $\alpha = 0.1$?

Soln:

Step 1. $H_0 : \mu = 1.0$, $H_1 : \mu \neq 1.0$.
Step 2. Use $\alpha = 0.05$
Step 3. $z_{0.025} = 1.96$
Step 4. $n = 12$, $\bar{x} = 1.2$, $\sigma = 0.4$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = 1.732 \qquad \text{vs.} \qquad 1.96$$

Step 5. H_0 cannot be rejected at $\alpha = 0.05$.
Step 6. p-value $= 2 \times P(z > 1.732) = 2 \times 0.0416 = 0.0832$
What happens at $\alpha = 0.1$?

Example 2.6. (Example 2.5 contd.) Suppose the sample standard deviation of metabolite X is 0.6 mg/dL. Assume that the (population) standard deviation of X is not known, and perform the hypothesis test.

Soln:

Step 1. $H_0 : \mu = 1.0$, $H_1 : \mu \neq 1.0$.
Step 2. Use $\alpha = 0.05$
Step 3. $t_{0.025,11} = 2.201$
Step 4. $n = 12$, $\bar{x} = 1.2$, $s = 0.6$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 1.155 \qquad \text{vs.} \qquad 2.201$$

Step 5. Cannot reject H_0.
Step 6. $P(t < 1.155)$ on a t_{11} curve is 0.8637.

$$p = 2 \times (1 - 0.8637) = 0.272$$

Example 2.7. The nationwide average score in an IQ test = 120. In a poor area, 100 student scores are analyzed and mean = 115, and standard deviation is 24. Is the area average lower than the national average?

Soln:

Step 1. $H_0 : \mu = 120$ vs. $H_1 : \mu < 120$

Step 2. Use $\alpha = 0.05$.

Step 3. $t_{0.05,99} = 1.66$. $z_{0.05} = 1.646$. Use either value.
Since we have $H_1 : \mu < 120$, our critical level is -1.66

Step 4.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{115 - 120}{24/\sqrt{100}} = -2.08 \qquad \text{vs.} \qquad -1.66$$

Step 5. We reject H_0 at a significance level of 0.05.

Step 6. The p-value is $P(t_{99} < -2.08) = 0.02$.

- Result is statistically significant.
- Could not have rejected H_0 at significance level 0.01.

Example 2.8. (Example 2.7 contd.) Suppose 10,000 scores were measured with mean = 119 and $s = 24$. Redo the test.

Soln:

Step 1. $H_0 : \mu = 120$ vs. $H_1 : \mu < 120$

Step 2. Use $\alpha = 0.05$.

Step 3. $t_{0.05,9999} = z_{0.05} = 1.646$.
Since we have $H_1 : \mu < 120$, our critical level is -1.646

Step 4.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{119 - 120}{24/\sqrt{10000}} = -4.17 \qquad \text{vs.} \qquad -1.646$$

Step 5. We reject H_0 at a significance level of 0.05.

Step 6. The p-value is $P(t_{9999} < -4.17) = P(z < -4.17) < 0.001$.

- Result is very statistically significant.
- Result is probably not scientifically significant.

Example 2.9. A pilot study of a new BP drug is performed for the purpose of planning a larger study. Five patients who have a mean BP of at least 95 mm Hg are recruited for the study and are kept on the drug for 1 month. After 1 month the observed mean decline in BP in these 5 patients is 4.8 mm Hg with a standard deviation of 9 mm Hg.

If μ_d= true mean difference in BP between baseline and 1 month, then how many patients would be needed to have a 90% chance of detecting a significant difference using a one-tailed test with a significance level of 5%?

Example 2.10. (Example 2.9 contd.) We go ahead and study the preceding hypothesis based on 20 people (we could not find 31). What is the probability that we will be able to reject H_0 using a one-sided test at the 5% level if the true mean and standard deviation of the BP difference is the same as that in the pilot study?

3. The χ^2-Test: One sample test for the variance of a Normal population.

Step 1. State the hypothesis.

$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 \neq \sigma_0^2.$

Standardize s^2 and use χ^2_{n-1} curve.

We need both thresholds: $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

Use $\alpha = 0.05$, on the $\chi^2_{\alpha/2,n-1}$ curve

Step 4. Compute the test statistic from the sampled data.

$$\frac{(n-1)s^2}{\sigma_0^2} \qquad \text{vs.} \qquad \chi^2_{n-1}$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

Example 2.11. We desire to estimate the concentration (μg/mL) of a specific dose of ampicillin in the urine after various periods of time. We recruit 25 volunteers who have received ampicillin and find that they have a mean concentration of 7.0 μg/mL, with a standard deviation of 2.0 μg/mL. Assume that the underlying distribution of concentrations is normally distributed.

1. Find a 95% confidence interval for the population mean concentration.

Soln: $(6.17, 7.83)$

2. Find a 99% confidence interval for the population variance of the concentrations.

Soln: $(2.11, 9.71)$

Example 2.12. (Example 2.11 contd.) How large a sample would be needed to ensure that the length of the confidence interval in part (a) (for the mean) is 0.5 μg/mL if we assume that the standard deviation remains at 2.0 μg/mL?

Soln: $n = 246$

Example 2.13. Iron deficiency causes anemia. 21 students from families below the poverty level were tested and the mean daily iron intake was found to be 12.50 mg with standard deviation 4.75 mg. For the general population, the standard deviation is 5.56 mg. Are the two groups comparable?

Soln: 14.6 lies within (9.591, 34.170) and hence cannot reject H_0.

Soln: Find the p-value yourself.

---

4. A Test for the Binomial Proportion

One sample test for p assuming that the Normal approximation to the Binomial is valid

Step 1. State the hypothesis.

$H_0 : p = p_0, \quad H_1 : p \neq p_0.$

Under H_0, $x \sim \mathcal{N}(np_0, np_0(1-p_0))$.

Standardize and use $\mathcal{N}(0,1)$ curve.

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

Use $\alpha = 0.05$, on the z curve

Step 4. Compute the test statistic from the sampled data.

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \qquad \text{vs.} \qquad z_{\alpha/2}$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

5. A Test for the equality of means of two Normal populations, when the variances are known

Consider measurements x_i and y_i arising out of two different populations. We assume that we have n measurements for x_i and m for y_i.

$$x_i \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right) \qquad y_i \sim \mathcal{N}\left(\mu_y, \sigma_y^2\right) \tag{2.104}$$

Then the point estimates of the means for the two populations are

$$\bar{x} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n}\right) \qquad \bar{y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma_y^2}{m}\right) \tag{2.105}$$

This therefore implies that

$$\bar{x} - \bar{y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right) \tag{2.106}$$

Remember that variances of independent variables add up.

Step 1. State the hypothesis.

$H_0 : \mu_x - \mu_y = 0, \quad H_1 : \mu_x - \mu_y \neq 0.$

Under H_0, $\bar{x} - \bar{y}$ follows Eq. 2.106. Standardize and use $\mathcal{N}(0,1)$ curve.

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

At $\alpha = 0.05$, on the Standard Normal, for a two-sided test, $z_{\alpha/2} = 1.96$.

Step 4. Compute the test statistic from the sampled data.

$$z = \frac{(\bar{x} - \bar{y}) - \cancelto{0}{(\mu_x - \mu_y)}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \qquad \text{vs.} \qquad \mathcal{N}(0,1)$$

If $\sigma_x^2 = \sigma_y^2 = \sigma^2$ (as would be expected when the samples sets are expected to be coming from the same population under H_0), then

$$\bar{x} - \bar{y} \sim \mathcal{N}\left(\mu_x - \mu_y, \sigma\left(\frac{1}{n} + \frac{1}{m}\right)\right) \tag{2.107}$$

The statistic therefore is

$$z = \frac{(\bar{x}-\bar{y}) - (\cancelto{0}{\mu_x - \mu_y})}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad \text{vs.} \qquad \mathcal{N}(0,1)$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

6. A Test for the equality of means of two Normal populations, when the variances are unknown but assumed to be equal

We have a situation where the population variances are unknown, but we assume that $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Then Eq. 2.107 does not hold and instead of the Gaussian, we need a t-distribution.

$$\bar{x}-\bar{y} \sim t\left(\mu_x - \mu_y, \sigma\left(\frac{1}{n}+\frac{1}{m}\right)\right) \tag{2.108}$$

For the individual populations we have the separate sample variances.

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad S_y^2 = \frac{\sum_{i=1}^{m}(y_i-\bar{y})^2}{m-1} \tag{2.109}$$

i.e. the estimate of σ^2 should not be unduly influenced by one of the sample sets.

We need to estimate σ^2 and that all $n+m$ data points must help towards this. The best estimate of σ^2 is the pooled sample variance, S_p^2, where

$$S_p^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{m}(y_i-\bar{y})^2}{n+m-2} = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2} \tag{2.110}$$

The relevant t distribution is one with $n+m-2$ degrees of freedom.

Step 1. State the hypothesis.

$H_0 : \mu_x - \mu_y = 0, \quad H_1 : \mu_x - \mu_y \neq 0.$

Under H_0, $\bar{x}-\bar{y}$ follows Eq. 2.108. Standardize and use t_{n+m-2} curve.

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

Use $\alpha = 0.05$, on the t_{n+m-2} curve

Step 4. Compute the test statistic from the sampled data.

$$t = \frac{(\bar{x}-\bar{y}) - (\cancelto{0}{\mu_x - \mu_y})}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad \text{vs.} \qquad t_{n+m-2}$$

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

7. A Test for the equality of means of two Normal populations, when the variances are unknown and found to be unequal

Satterthwaite's method. Also look up the 'Behrens-Fisher' problem.

Eq. 2.106 cannot be simplified to Eq. 2.108 because the population variances are unequal, and correspondingly, a pooled variance cannot be used. The test statistic now becomes (under H_0)

$$t = \frac{(\bar{x}-\bar{y}) - (\cancelto{0}{\mu_x - \mu_y})}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \tag{2.111}$$

The difficulty that remains is identifying the degrees of freedom of the t distribution that needs to be looked up. The recommended approach is to find the degrees of freedom d'' as follows:

$$d' = \frac{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1}+\frac{\left(S_y^2/m\right)^2}{m-1}} \tag{2.112}$$

$$d'' = d' \text{ rounded down} \tag{2.113}$$

Step 1. State the hypothesis.
$H_0 : \mu_x - \mu_y = 0, \quad H_1 : \mu_x - \mu_y \neq 0.$
Under H_0, $\bar{x} - \bar{y}$ follows Eq. 2.106. Standardize and use $t_{d''}$ curve.

Step 2. Identify the significance level (α) at this stage.
We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).
Use $\alpha = 0.05$, on the $t_{d''}$ curve

Step 4. Compute the test statistic from the sampled data using Eq. 2.111.

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

8. A Test for the equality of variances of two Normal populations (F-test)

When the population variances for the two Normal populations are unknown, we have (in the previous two tests) developed tests for equality of the means, assuming that the variances are either equal (Case 6) or unequal (Case 7). What we should have been doing, before testing for the equality of the means, was to first ask whether the variances could be assumed equal, especially considering that the methodology in Case 6 is simpler than that in Case 7. From the discussion on the χ^2 distribution, we have a ratio of two χ^2 distributions (with in general different degrees of freedom), which is a new, F, distribution, dependent on the two sets of degrees of freedoms.

$$\frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi^2_{n-1} \qquad \frac{(m-1)S_y^2}{\sigma_y^2} \sim \chi^2_{m-1} \tag{2.114}$$

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} = F_{n-1,m-1} \tag{2.115}$$

Under $H_0 : \sigma_x^2 = \sigma_y^2$, we get

$$\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1} \tag{2.116}$$

Then H_0 cannot be rejected if

$$F_{1-\alpha/2,n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq F_{\alpha/2,n-1.m-1} \tag{2.117}$$

Note that the outcome of the F-test should not depend on which data set is in the numerator of Eq. 2.116.

The choice of which set is labeled x and y is not important given the following equivalence between two F distributions:

$$P\left[\frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} > F_{\alpha,n-1,m-1}\right] = \alpha \quad \Rightarrow \quad P\left[\frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} < \frac{1}{F_{\alpha,n-1,m-1}}\right] = \alpha \tag{2.118}$$

This implies that

$$P\left[\frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} > \frac{1}{F_{\alpha,n-1,m-1}}\right] = 1-\alpha = P\left[\frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} > F_{1-\alpha,m-1,n-1}\right] \tag{2.119}$$

which shows that

$$\frac{1}{F_{\alpha,n-1,m-1}} = F_{1-\alpha,m-1,n-1} \tag{2.120}$$

The F-test is summarized as follows:

Step 1. State the hypothesis.
$H_0 : \sigma_x^2 = \sigma_y^2, \quad H_1 : \sigma_x^2 \neq \sigma_y^2.$
Under H_0, we expect Eq. 2.116.

Step 2. Identify the significance level (α) at this stage.
We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).
Use $\alpha = 0.05$, on the $F_{n-1,m-1}$ curve

Step 4. Compute the test statistic, S_x^2/S_y^2, from the sampled data.

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

Sample sizes for two sample tests of the means

Consider $H_0 : \mu_x = \mu_y$ and $H_1 : \mu_x \neq \mu_y$. Under H_0, the lower critical threshold is (in standardized coordinates) $-z_{\alpha/2}$. In the original coordinate system, relative to the H_0 curve, this threshold is

$$\cancelto{0}{(\mu_x - \mu_y)} - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} = -z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

But this coordinate, relative to the H_1 curve is known to be (after standardization), z_β.

$$z_\beta = \frac{\left(-z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}\right) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \tag{2.121}$$

This can be rearranged to get, using $k = m/n$,

$$\left(\frac{z_{\alpha/2} + z_\beta}{\mu_x - \mu_y}\right)^2 = \frac{n}{\sigma_x^2 + \sigma_y^2/k} \tag{2.122}$$

and hence, given α and β, the sample sizes required to differentiate between two populations are

$$n = \left(\frac{z_{\alpha/2} + z_\beta}{\mu_x - \mu_y}\right)^2 \left(\sigma_x^2 + \frac{\sigma_y^2}{k}\right) \tag{2.123}$$

$$m = nk \tag{2.124}$$

9. A Test for the equality of means of two Normal populations, when the samples are paired (the paired t-test)

Consider n pairs of before and after measurements, which reflect, for example, the efficacy of some treatment. Then the n differences d_i can be used to compute the mean sample difference on treatment ($= \bar{d}$), and the sample standard deviation $sd(d_i)$. We have one vector of difference values, with the average difference expected to be zero, according to H_0. Since the population variation is unknown, we have a t test with $n-1$ degrees of freedom.

$$\frac{\bar{d} - \mathcal{E}\left[\bar{d}\right]}{sd(\bar{d})} = \frac{\bar{d} - \cancelto{0}{\mathcal{E}\left[\bar{d}\right]}}{sd(d_i)/\sqrt{n}} \sim t_{n-1} \tag{2.125}$$

Then, the paired t-test may be summarized as follows.

Step 1. State the hypothesis.
$H_0 : \mathcal{E}\left[\bar{d}\right] = \Delta = 0, \quad H_1 : \Delta \neq 0.$

Under H_0, we expect Eq. 2.125 to hold.

Step 2. Identify the significance level (α) at this stage.

We set $\alpha = 0.05$.

Step 3. Based on H_1 and α identify the critical threshold(s).

Use $\alpha = 0.05$, on the t_{n-1} curve

Step 4. Compute the test statistic, $\sqrt{n}\bar{d}/sd(d_i)$, from the sampled data.

Step 5. Identify the test result given the critical level (based on α).

Step 6. Report the test result, and the p-value.

2.5 Nonparametric hypothesis testing

When the data available to us

- does not in any fashion obviously follow some distribution,
- or when the data is ordinal in nature rather than cardinal,

the parametric hypothesis tests previously described cannot be applied. We then resort to nonparametric hypothesis tests. Two essential features of these tests are that we pay attention to the median of a data set, rather than the arithmetic mean, and that we endeavor to identify a statistic, like a proportion, or a rank, which can be expected to follow a parametric distribution. In this manner, we drag things back towards the more convenient parametric hypothesis tests. We review below three nonparametric tests which will be seen to be counterparts of some of the two sample parametric tests already discussed.

1. The Wilcoxon sign test.
2. The Wilcoxon signed rank test.
3. The Wilcoxon rank sum test.

The Wilcoxon sign test

The paired t-test is applied under the assumption that all measurements follow some Normal distribution. If the underlying distribution is not Normal, or if it is unknown, then this test is a safer test.

The Wilcoxon sign test is the nonparametric counterpart to the paired t test, and is usually invoked to evaluate whether two data sets are similar. Let d_i indicate the relative like or dislike for two categories (icecreams). Then we have 3 categories of data: $d_i < 0$ where flavor A is preferred, $d_i > 0$ with B being preferred, and $d_i = 0$ with equal preference for both flavors. We wish to determine whether there is a marked preference for one flavor versus the other. H_0 would suggest an equal number of people preferring the two icecreams. The actual numbers in the survey voting for the two icecreams should turn out to be close. Note that those who cannot make up their minds, with $d_i = 0$, are irrelevant to the test of H_0. Therefore, if n scores have $d_i \neq 0$, of which c have $d_i > 0$, the test itself will be to see if c is close enough to the median of the data set, $n/2$. If $\Delta = \mathcal{E}[c]$ is the population median of the underlying distribution of d_i, then

$$H_0 : \mathcal{E}[c] = \Delta = \frac{n}{2} \qquad H_1 : \Delta \neq \frac{n}{2}$$

Under H_0, the binomial proportion of finding nonzero $d_i > 0$ is $1/2$. Further, assuming the Binomial → Normal approximation, $\text{Var}[c] = np(1-p) \geq 10$, and under H_0, since $p = 1/2$, we get $n \geq 40$. Then we reject H_0 if

$$\left| \frac{c - \mathcal{E}[c]}{sd(c)} \right| = \left| \frac{c - n/2}{\sqrt{n/4}} \right| > z_{\alpha/2} \tag{2.126}$$

A continuity correction is required because we are shifting over from a discrete to a continuous distribution. H_0 is rejected if

$$\begin{aligned} \text{If } c > \frac{n}{2}: \qquad & \left| \frac{c - \frac{n}{2} - \frac{1}{2}}{\sqrt{n/4}} \right| > z_{\alpha/2} \\ \text{If } c < \frac{n}{2}: \qquad & \left| \frac{c - \frac{n}{2} + \frac{1}{2}}{\sqrt{n/4}} \right| > z_{\alpha/2} \end{aligned} \tag{2.127}$$

Example 2.14. 2 flavours of icecream undergo a taste test. 45 people are asked to choose, and 22 prefer A, 18 prefer B, and 5 cannot make up their minds. Is one flavour preferred over the other?

Soln:

$n = 40$, $c = 18$, the statistic is

$$\frac{(18 - 20) + (1/2)}{\sqrt{40/4}} = \frac{-1.5}{\sqrt{10}} = -0.47$$

The p-value is $2 \times (1 - \Phi(0.47)) = 2 \times (1 - 0.6808) \simeq 0.64$. Hence H_0 cannot be rejected according to this test.

The Wilcoxon signed rank test

Rather than simply resort to proportions employed in the Sign test, we bring in some more detail into the hypothesis test, but ordering the measurements and assigning ranks to them. We then expect the ranks to be equally distributed between corresponding quantile groups on either side of the median. This therefore ends up testing whether the distribution function has similar shapes on either side of the median, as would be expected of H_0 where the two groups are supposed to be identical. The procedure is as follows:

1. Arrange the d_i's in order of absolute value and get frequency counts.
2. As in the sign test, ignore the $d_i = 0$ observations.
3. Rank the remaining observations from 1 to n.
4. Let R_1 be the sum of ranks for group 1.
5. H_0 assumes that the ranks are equally shared between the two groups, and H_1 assumes otherwise. Since n ranks are given out, the sum of all ranks is $n(n+1)/2$, and hence the rank sum for group 1 should be $n(n+1)/4$.
$$\mathcal{E}\left[R_1\right] = \frac{n(n+1)}{4}, \qquad \text{Var}\left[R_1\right] = \frac{n(n+1)(2n+1)}{24} \tag{2.128}$$
6. For $n \geq 16$, we assume that the rank sum follows a Normal distribution
$$R_1 \sim \mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) \tag{2.129}$$
7. The test statistic T, bringing in the continuity correction again, is
$$\frac{|R_1 - \frac{n(n+1)}{4}| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{2.130}$$
8. A further correction factor is required for the variance, if there are subgroups with tied

ranks. The variance calculated above will be an overestimate. If g indicates the number of tied groups, and t_i the number of tied values in that group, then the corrected variance is

$$\mathrm{Var}\left[R_1\right] = \frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{g} \frac{t_i^3 - t_i}{48} \qquad (2.131)$$

Equation 2.130 needs to be correspondingly updated.

	$d_i < 0$	$d_i > 0$
$\lvert d_i \rvert$	f_i	f_i
10	0	0
9	0	0
8	1	0
7	3	0
6	2	0
5	2	0
4	1	0
3	5	2
2	4	6
1	4	10

Example 2.15. (Example 2.14 contd.) The 45 volunteers rate each icecream on a scale from 1 to 10. (data in the margin). (10 = really really like it, 1 = don't care). Let d_i = score(B) - score(A). Then $d_i > 0 \Rightarrow$ B is preferred.

Soln:

	$d_i < 0$	$d_i > 0$	Total # of	Range of	Average
$\lvert d_i \rvert$	f_i	f_i	people	ranks	rank
10	0	0	0		
9	0	0	0		
8	1	0	1	40	40
7	3	0	3	37-39	38
6	2	0	2	35-36	35.5
5	2	0	2	33-34	33.5
4	1	0	1	32	32
3	5	2	7	25-31	28.0
2	4	6	10	15-24	19.5
1	4	10	14	1-14	7.5

We conveniently choose R_1 to be the group corresponding to $d_i > 0$ since it has fewer nonzero frequencies (the arithmetic becomes simpler).

$$R_1 = 10(7.5) + 6(19.5) + 2(28.0) = 248$$

$$\mathcal{E}\left[R_1\right] = \frac{40 \times 41}{4} = 410$$

$$\mathrm{Var}\left[R_1\right] = \frac{40 \times 41 \times 81}{24} - \frac{(14^3 - 14) + (10^3 - 10) + (7^3 - 7) + (3^3 - 3) + 2 \times (2^3 - 1)}{48} = 5449.75$$

$$sd(R_1) = \sqrt{5448.75} = 73.82$$

$$T = \frac{|(248 - 410)| - (1/2)}{73.82} = 2.19$$

The p-values $= 2(1 - 0.9857) \simeq 0.03$ and hence H_0 is rejected. Notice that by including the quantile level information, the signed rank test gives a different conclusion to that of the Sign test. Since this test has involved more detailed analysis, the conclusion here is preferable to that in Example 2.14. Observe, from the method employed, that this test is the nonparametric equivalent of the paired t test.

Likely because flavor A is preferred.

The Wilcoxon rank sum test

aka Mann Whitney test aka U Test

The rank sum test is the nonparametric equivalent of the t-test for 2 independent samples. The procedure involves assigning ranks after combining the data from the two groups. As with the signed rank test, the rank sum for a group is compared with the expected rank sum. Let R_1 represent the rank sum in the first group, and let there be n_1 and n_2 samples in the two groups. Then, the rank sum of all the ranks is $(n_1+n_2)(n_1+n_2+1)/2$ and hence the average rank for the combined sample is $(n_1+n_2+1)/2$. Under H_0,

$$\mathcal{E}\left[R_1\right] = n_1\frac{n_1+n_2+1}{2} \qquad \operatorname{Var}\left[R_1\right] = n_1 n_2\frac{n_1+n_2+1}{12} \tag{2.132}$$

If n_1 and n_2 are ≥ 10, the rank sum may be assumed to have an underlying continuous (Normal) distribution:

$$R_1 \sim \mathcal{N}\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right) \tag{2.133}$$

and hence the appropriate test statistic is

$$T = \frac{|R_1 - \frac{n_1(n_1+n_2+1)}{2}| - \frac{1}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \tag{2.134}$$

Once again a correction factor needs to be applied to correct for groups of tied values. If g indicates the number of tied groups, and t_i the number of tied values in that group, then the corrected variance is

$$\operatorname{Var}\left[R_1\right] = \frac{n_1 n_2(n_1+n_2+1-S)}{12} \qquad S = \sum_{i=1}^{g}\frac{t_i^3 - t_i}{(n_1+n_2)(n_1+n_2-1)} \tag{2.135}$$

Equation 2.134 needs to be correspondingly updated.

	Group M	Group N
a	5	1
b	9	5
c	6	4
d	3	4
e	2	8
f	0	5
g	0	2
h	0	1

Example 2.16. Consider the frequencies observed for two groups (data in the margin). The first column could represent grade labels for instance.

Soln:

	Group M	Group N freq	Total range	rank rank	average
a	5	1	6	1-6	3.5
b	9	5	14	7-20	13.5
c	6	4	10	21-30	25.5
d	3	4	7	31-37	34.0
e	2	8	10	38-47	42.0
f	0	5	5	48-52	50.0
g	0	2	2	53-54	53.5
h	0	1	1	55	55.0

$$R_1 = 5(3.5) + 9(13.5) + 6(25.5) + 3(34) + 2(42.5) = 479$$

$$\mathcal{E}\left[R_1\right] = \frac{25\times 26}{2} = 700$$

$$A = (6^3-6) + (14^3-14) + (10^3-10) + (7^3-7) + (10^3-10) + (5^3-5) + (2^3-2) = 5387$$

$$\operatorname{Var}\left[R_1\right] = \frac{25\times 30}{12}\left(56 - \frac{A}{55\times 54}\right) = 3386.75$$

$$T = \frac{|479 - 700| - 0.5}{\sqrt{3386.74}} = 3.79$$

Clearly 3.79 is $> z_{\alpha/2}$ for the common choices of α, and the p values will be < 0.001. Hence H_0 is rejected and the groups are significantly different.

2.6 Exercises

1. A clinical trial is conducted to evaluate the efficacy of spectinomycin, as a treatment for a disease. 46 patients are given a 4-g daily dose of the drug and are seen one week later, at which time 6 patients still have the disease.
    1. What is the best point estimate for p, the probability of a failure with the drug?

    Soln: 6/46

    2. What is a 95% confidence interval for p?

    Soln: (0.033, 0.228)

    3. Suppose that we know that penicillin G at a daily dose of 4.8 mega units has a 10% failure rate. What can be said in comparing the two drugs?
2. The fraction of defective microprocessors chips produced by a new process is being studied. A random sample of 300 chips is tested, revealing 13 defectives.
    1. Find a 95% two-sided confidence interval on the fraction of defective chips produced by this process.
    2. Do the data support the claim that the fraction of defective units produced is **less than** 0.05, using $\alpha = 0.05$?

    Soln: H_0 cannot be rejected.

    3. Find the P-value for the test.
3. The two sample t test: testing for for equality of means of two sets of independent normally distributed samples, with variances unknown but assumed equal. 8 BP drug users have a mean BP of 132.86 mm Hg and a sample SD = 15.34 mm Hg. 21 non-drug users have a mean BP of 127.44 and a sample SD of 18.23 mm. Does the drug have any effect?
4. It is believed that cigarette smoking at home affects the lung function of children. Suppose two groups of children aged 5-9 are studied and the lung FEV measured:

Group 1 23 nonsmoking children; *both* parents smoke; FEV: mean of 2.1 L and sd of 0.7 L.

Group 2 20 nonsmoking children; *neither* of whose parents smoke; FEV: mean of 2.3 L and sd of 0.4 L.

    1. What are the appropriate null and alternative hypotheses?
    2. What is the appropriate test procedure for (1.)?
    3. Carry out the test in (2.) using the critical-value method.
    4. Provide a 95% confidence interval for the true mean difference in FEV between 5- to 9-year-old children whose parents smoke and comparable children whose parents do not smoke.
    5. If this is regarded as a pilot study, then how many children are needed in each group (assuming equal numbers in each group) to have a 95% chance of detecting a significant difference using a two-sided test with $\alpha = 0.05$?
    6. Repeat (5.) using a one-sided test.

7. Suppose 40 children, both of whose parents smoke, and 50 children, neither of whose parents smoke, are recruited for the study. How much power would such a study have using a two-sided test with significance level = 0.05, assuming that the estimates of the population parameters in the pilot study are correct?
8. Repeat (7.) using a one-sided test.

5. The mean ±1 standard deviation of log[calcium intake in mg] among 25 12- to 14-year-old females below the poverty line is 6.56 ± 0.64. Similarly, for 40 12- to 14-year-old females above the poverty line, mean ±1 standard deviation is 6.80 ± 0.76.

   1. Suppose an equal number of 12- to 14-year-old girls below and above the poverty line are recruited to study differences in calcium intake. How many girls should be recruited to have an 80% chance of detecting a significant difference using a two-sided test with $\alpha = 0.05$?
   2. Repeat (1.) using a one-sided rather than a two-sided test.
   3. Using a two-sided test with $\alpha = 0.05$, solve (1.) anticipating that 2 girls above the poverty level will be recruited for every girl below the poverty level.
   4. Suppose 50 girls above the poverty level and 50 girls below the poverty level are recruited for the study. How much power will the study have of finding a significant difference using a two-sided test with $\alpha = 0.05$ assuming that the population parameters are the same as the sample estimates in (1.)?
   5. Repeat (4.) using a one-sided test rather than a two-sided test.
   6. Suppose that 50 girls above the poverty level and 25 girls below the poverty level are recruited for the study. How much power will the study have if a two-sided test is used with $\alpha = 0.05$?
   7. Repeat (6.) if a one-sided test is used with $\alpha = 0.05$.

6. One method for assessing the effectiveness of a drug is to note its concentration in blood and/or urine samples at certain periods of time after giving the drug. Suppose we wish to compare the concentrations of two types of aspirin (types A and B) in urine specimens taken from the same person, 1 hour after he or she has taken the drug. Hence, a specific dosage of either type A or type B aspirin is given at one time and the 1-hour urine concentration is measured. One week later, after the first aspirin has presumably been cleared from the system, the same dosage of the other aspirin is given to the same person and the 1-hour urine concentration is noted. Because the order of the drugs may affect the results, a table of random numbers is used to decide which of the two types of aspirin to give first. This experiment is performed on 10 people; the results are given below. (Aspirin A and Aspirin B values in the table are the 1-hour concentrations in mg%).

Person	Aspirin A	Aspirin B
1	15	13
2	26	20
3	13	10
4	28	21
5	17	17
6	20	22
7	7	5
8	36	30
9	12	7
10	18	11
Mean	19.20	15.60
sd	8.63	7.78

Suppose we wish to test the hypothesis that the concentrations of the two drugs are the same in urine specimens.

1. What are the appropriate hypotheses?
2. What are the appropriate procedures to test these hypotheses?
3. Conduct the tests.
4. What is the best point estimate of the mean difference in concentrations between the two drugs?
5. What is a 95% confidence interval for the mean difference?

7. Diflunisal is a drug used to treat arthritis. Its effects on intraocular pressure in glaucoma patients who were already receiving maximum therapy for glaucoma, was conducted.

   1. Suppose the change (mean ± sd) in ocular pressure after administration of Diflunisal (follow-up - baseline) among 10 patients whose standard therapy was methazolamide and topical glaucoma medications was -1.6 ±1.5 mm Hg. Assess the statistical significance of the results.
   2. The change in ocular pressure after administration of Diflunisal among 30 patients whose standard therapy was topical drugs only was -0.7 ± 2.1 mm Hg. Assess the statistical significance of these results.
   3. Find 95% confidence limits for the mean change in pressure in each of the two groups identified in (a) and (b).
   4. Compare the mean change in ocular pressure in the two groups identified in (a) and (b) using hypothesis-testing methods.

8. A big study was conducted of genetic and environmental influences on cholesterol levels. The full data set used involved four populations of adult twins:

   1. monozygotic (MZ) twins reared apart,
   2. MZ twins reared together,
   3. dizygotic (DZ) twins reared apart, and
   4. DZ twins reared together.

   Before the results are reported, an issue is whether it is necessary to correct for sex before commenting on genetic influences.

   The data below is for total cholesterol levels for MZ twins reared apart, categorized by sex. (n = number of people: for males, 22 pairs of twins = 44 people).

	men	women
Mean	253.3	271.0
sd	44.1	44.1
n	44	48

   1. If we assume that (1) serum cholesterol is normally distributed, and (2) the samples are independent, and (3) the standard deviations for men and women are the same, then which statistical procedure can be used to compare the two groups?
   2. State the hypothesis to be tested by the procedure in (a). Use a two-sided test and report a p-value.
   3. Use the procedure in (a) with a one-sided test where the alternative hypothesis is that men have higher cholesterol levels than women. State the hypotheses being tested and implement the method in (a). Report a p-value.
   4. Are the assumptions in (a) likely to hold for these samples? Why or why not?

9. The mean systolic blood pressure (SBP) of 20 patients is measured with two different machines (A and B), and the data is presented below.

Person	1	2	3	4	5	6	7	8	9	10
A	79	112	103	104	94	106	103	97	88	113
B	84	99	92	103	94	106	97	108	77	94

Person	11	12	13	14	15	16	17	18	19	20
A	98	103	105	117	94	88	101	98	91	105
B	97	103	107	120	94	87	97	93	87	104

1. Which test should be used to test the hypothesis that the mean SBP for the two machines is comparable?

   Soln: Wilcoxon signed rank test (large sample test)
2. Conduct the test.

   Soln: $R_1 = 33.5$, $T = 1.76$ vs. $N(0, 1)$, $p = 0.078$

10. A clinical trial was undertaken comparing PTCA (a type of angioplasty) with medical therapy in the treatment of single-vessel coronary-artery disease. 107 patients were randomly assigned to medical therapy and 105 to angioplasty (PTCA). Patients were given exercise tests at baseline and after 6 months of follow-up. Exercise tests were performed up to maximal effort until symptoms such as angina were present. The results shown below were obtained for change in total duration of exercise (min) (6 months - baseline).

	Mean change (min)	sd	n
medical therapy	0.5	2.2	100
PTCA	2.1	3.1	99

1. What test can be performed to see if there has been a change in mean total duration of exercise for a specific treatment group?
2. Perform the test in (a) for the medical therapy group, and report a p-value.
3. What test can be performed to compare the mean change in the duration of exercise *between* the two treatment groups.
4. Perform the test mentioned in (c) and report a p-value.

11. Suppose we wish to compare the length of stay in the hospital for patients with the same diagnosis at two different hospitals. The following results are found:

First hospital	21, 10, 32, 60, 8, 44, 29, 5, 13, 26, 33
Second hospital	86, 27, 10, 68, 87, 76, 125, 60, 35, 73, 96, 44, 238

1. Why might a t test not be very useful in this case?
2. Carry out a nonparametric procedure for testing the hypothesis that the lengths of stay are comparable in the two hospitals.