STA1000F
Summary

Mitch Myburgh
MYBMIT001

May 28, 2015

1 Module 1: Probability

1.1 Work Unit 1: Introducing Probability

1.1.1 Definitions

1. **Random Experiment:** A procedure whose outcome (result) in a particular performance (trial) cannot be predetermined.
2. **Long Run Pattern (Average):** the average result over a large number of trials
3. **Random Variation:** This implies that we never know the value of the next random event
4. **Statistical Distributions:** Distributions of data
5. **Fair Game:** No one wins or loses in the long run
6. **House Advantage:** The profit made by the house (casino)

1.1.2 Formula

1.
$$Win\% = \frac{Total\ Payout\ for\ a\ winning\ number}{amount\ to\ be\ bet\ over\ all\ numbers} \times 100$$

2.
$$Win\% \times fair\ payout = payout$$

3.
$$fair\ payout - fp \times HA = payout$$

4.
$$fair\ payout = (probability\ of\ winning)^{-1} \times bet$$

5.
$$HA = 100 - Win\%$$

1.1.3 Examples

1. Die: S = {1,2,3,4,5,6} P(6) = ⅙ P(even) = 3⁄6 = ½
2. Odds: odds of {6} = 1:5

1.2 Work Unit 2: Set Theorem, Probability Axioms and Theorems

1.2.1 Definitions

1. Sets can be determined by a list of elements ($A = \{e, f, g, 1, 2\}$) or a rule ($B = \{x|1 \leq x \leq 10,\ x \in \mathbb{Z}\}$)
2. Order and repetition in sets is irrelevant $\{1, 3, 4, 4, 2\} = \{1, 2, 3, 4\}$
3. Element member of set ($e \in A$) vs Element not member ($e \notin B$)
4. **Subsets:** if $G \subset H$ and $G \supset H$ then $G = H$

5. **Intersection:** $A \cap B = \{1, 2\}$

6. **Union:** $A \cup B = \{e, f, g, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

7. **Complement:** $S = \{1, 2, 3, 4\}$, $C = \{1, 2\}$, $\bar{C} = \{3, 4\}$

8. **Empty Set:** $\emptyset = \{\}$

9. **Universal Set (Sample Sapce)** S - all possible outcomes of a random experiment

10. **Mutually Exclusive (Disjoint):** If $L \cap M = \emptyset$

11. **Pairwise Mutually Exclusive, Exhaustive Sets:** $A_1, A_2, \ldots, A_n$ s.t. $A_i \cap A_j = \emptyset$ if $i \neq j$ and $A_1 \cup A_2 \cup \ldots \cup A_n = S$

12. **Event:** Subset of sample space (S = certain event, $\emptyset$ = impossible event)

13. **Elementary Event:** Event with one member ($A = \{3\}$) Always mutually exclusive, $P(A) = {}^{n(A)}/_{n(S)}$. NB not $\emptyset$

14. A occurs if the outcome of the trial is a member of A

15. **Relative Frequency:** r⁄n, r= number of times A occurs, n = number of trials, $0 \leq {}^{r}/_{n} \leq 1$, $P(A) = \lim_{n \to \infty} {}^{r}/_{n}$

1.2.2 Formula

1. $$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

2. $$A \cup B = A \cup (B \cap \bar{A})$$

3. Kolmogorov's Axioms of Probability
   S = sample space, $\forall A \subset S$, $P(A) \in \mathbb{R}$ st

   (a) $0 \leq P(A) \leq 1$

(b) $P(S) = 1$
(c) If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
(d) Consequence: $P(\emptyset) = 0$

4.
$$\text{Let } A \subset S \text{ then } P(\bar{A}) = 1 - P(A)$$

5.
$$\text{If } A \subset S \text{ and } B \subset S \text{ then } P(A) = P(A \cap B) + P(A \cap \bar{B})$$

6.
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

7.
$$\text{If } B \subset A \text{ then } P(B) \leq P(A)$$

8. If $\mathrm{A}_1, \ldots, A_n$ are pairwise mutually exclusive then: $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$

1.3 Work Unit 3: Permutations and Combinations

1.3.1 Counting Rules

1. Permutation: order matters, repetition not allowed
$$n!$$
2. Permutation: order matters, repetition not allowed
$$(n)_r = \frac{n!}{(n-r)!}$$
3. Permutations: order matters, repetition allowed
$$n^r$$
4. Combinations: Order doesn't matter, repetition not allowed
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

1.4 Work Unit 4: Conditional Probability and Independent Events

1. Conditional Probability:
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2. Complement of Conditional Probability:
$$P(P|D) = 1 - P(\bar{P}|D)$$

3. Baye's Theorem:
$$P(D|P) = \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|\bar{D})P(\bar{D})}$$

4. Baye's Theorem for mutually exclusive exhaustive events:
$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$

5. Independent Events: (never mutually exclusive) A and B independent
$$P(A|B) = P(A)$$
$$P(A \cap B) = P(A)P(B)$$
$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \times \ldots \times P(A_n)$$

2 Module 2: Exploring Data

2.1 Work Unit 1: Graphical Summaries

- Pie Charts: Categories add up to 100%; Units comparable; Slice reflect proportion; order biggest to smallest
- Bar Charts: order biggest to smallest
- Histograms: chose a number between ½L and 2L; all class intervals must be the same, Bimodal - has 2 peaks

2.1.1 Definitions

1. **Qualitative Data (Catagorical or Nominal Data):** e.g. Nationality/hair colour (use Pie Chart or Bar Chart)
2. **Quantitative Data (Fully Numeric):** Count/measure captured on scale (no fixed 0) or a ratio (explicit 0) e.g. Weight/height/distance (Use histogram, scatter plot, box plot)
3. **Ordinal Data:** Ranked or ordered data, steps don't have to be the same size, e.g. level of satisfaction/education

2.1.2 Formula

1. Interval Width

$$L = \frac{X_{max} - X_{min}}{\sqrt{n}}$$

   or

$$L = \frac{X_{max} - X_{min}}{1 + log_2 n} = \frac{X_{max} - X_{min}}{1 + 1.44 log_e n}$$

2. Trendline (linear regression)

$$y = a + bx$$

   a = intercept, b = slope

3. Explainatory Value R^2 The amount of variation in Y that can be explained by X

2.2 Work Unit 2: Summary Measures of Location and Spread

2.2.1 Definitions

1. **Statistic:** quantity calculated from the data values of a sample (subset of population)
2. **Parameter:** statistic calculated on the population

3. **5 Number Summary:** min; lower quartile; median; upper quartile; max

4. **Fences:** largest and smallest observations that aren't strays; whiskers in box-and-whisker plot go up to these when you also show outliers and strays

2.2.2 Measures of Location

1. **Median:** Robust; not affected by big outliers/strays

2. **Mean:** sensitive to outlying values; useful for symmetric distributions

2.2.3 Measures of Spread

1. **Range:** Unreliable/sensitve

2. **IQR:** Robust

3. **Standard Deviation:** $(\bar{x} - s, \bar{x} + s)$ contains ⅔ of your observations

2.2.4 Formula

1. Standard Deviation $= \sqrt{Variance}$

2. Variance

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

3. Update Variance:

$$s* = \sqrt{\frac{1}{n}[(n-1)s^2 + n(\bar{x} - \bar{x}*)^2 + (x_{n+1} - \bar{x}*)^2]}$$

4. Mean
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

5. Update Mean:
$$\bar{x}* = \frac{n\bar{x} + x_{n+1}}{n+1}$$

6. Strays
$$< Median - 3 \times (Median - LQ)$$
$$> Median + 3 \times (UQ - Median)$$

7. Outliers
$$< Median - 6 \times (Median - LQ)$$
$$> Median + 6 \times (UQ - Median)$$

8. Median = $X_{(m)}$ where $m = {}^{n+1}\!/_2$

9. LQ = $X_{(l)}$ where $l = {}^{[m]+1}\!/_2$

10. UQ = $X_{(u)}$ where $u = n - l + 1$

11. Range = $X_{max} - X_{min} = X_n - X_1$

12. IQR = $x_{(u)} - x_{(l)}$

13. Half Rank: $X_{(r+1/2)} = {}^{x_{(r)} + x_{(r+1)}}\!/_2$

2.3 Module 3: Random Variables

2.4 Work Unit 1: Probability Mass and Density Functions

- $P(X = x) = P(x)$, where x is specific value of random variable X
- you can assign numbers to qualitative events

2.4.1 Definitions

1. **Discrete Random Variable:** set of possible values is finite or countably infinite
   (a) **Probability Mass Function:** p(x)
   (b) Defined for all values of x, but non-zero at finite (or countably infinite) subset of these values
   (c) $0 \leq p(x) \leq 1 \; \forall x$
   (d) $\sum p(x) = 1$
   (e) $P(a \leq x < b) = \sum_{x=a}^{b-1} p(x)$
2. **Continuous Random Variable:** $\{x|a < x < b\}$
   (a) **Probability Density Function:** f(x)
   (b) Defined for all values of x
   (c) $0 \leq f(x) \leq \infty \; \forall x$
   (d) $\int_{-\infty}^{\infty} p(x) = 1$ check on non zero interval only
   (e) $P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) = P(a < X < b) = \int_a^b p(x)$
   (f) $P(X = a) = 0$
   (g) can be measured to any degree of accuracy

2.5 Work Unit 2: Working with Random Variables

2.5.1 Definitions

1. **Long Run Average:** Expected value of random variable X (E(X)), weighted sum of possible values of X, also called mean, can have theoretical value (value not in S e.g. 3.5 is E(X) for dice)
2. $\mu = E(X)$ and $\sigma^2 = Var(X)$

3. **Discrete Random Variable**

   (a) $E(X) = \sum xp(x)$

   (b) $Var(X) = \sum(x - E(X))^2 p(x) = (\sum x^2 p(x)) - E(X)^2$

   (c) $E(X^r) = \sum x^r p(x)$

4. **Continuous Random Variable**

   (a) $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

   (b) $Var(X) = E(X^2) - E(X)^2 = \int_a^b (x - E(X))^2 f(x)dx = (\int_a^b x^2 f(x)dx) - E(X)^2$

   (c) $E(X^r) = \int_a^b x^r f(x)dx$

5. **Expected Value:**

   (a) $E(A + B) = E(A) + E(B)$

   (b) $E(A - B) = E(A) - E(B)$

   (c) $E(cA) = cE(A)$ where c is constant

   (d) $Y = aX + b$ then $E(Y) = aE(X) + b$

6. **Variance:**

   (a) $Var(A + B) = Var(A) + Var(B)$

   (b) $Var(A - B) = Var(A) + Var(B)$

   (c) $Var(cA) = c^2 Var(A)$ where c is constant

   (d) $Y = aX + b$ then $Var(Y) = a^2 Var(X)$

7. **Coefficient of Variation (CV) =**

$$\frac{\sqrt{Var(X)}}{E(X)}$$

   only if lower limit of X is 0

8. in graphs of pdf, pmf peaked = small var, flat = large var
9. -vely skewed peak on right, symmetric peak in middle, +vely skewed peak on left
10. **Heavy-tailed Distributions:** Probabilty of observations far from the mean is relatively large
11. **Light-tailed distributions:** observations far from the mean are unlikely

2.6 Module 4: Probability Distributions

2.7 Work Unit 1: Uniform Distribution

$$X \sim U(a, b)$$

PDF:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & elsewhere \end{cases}$$

Integrate this to get the probability

Use this when you have a continuous random variable that is equally likely to lie between a and b and impossible to lie outside this interval.

$$E(X) = \frac{1}{2}(b + a)$$

$$Var(X) = \frac{(b-a)^2}{12}$$

Distribution Function:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

2.8 Work Unit 2: Binomial Distribution

$$X \sim B(n, p)$$

PMF:

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2...n \\ 0 & elsewhere \end{cases}$$

If we are observing the number of successes in n (fixed number) independent trials of an experiment in which the outcome of each trial can only be success or failure with constant probability

$$E(X) = np$$

$$Var(X) = np(1-p)$$

2.9 Work Unit 3: Poisson and Exponential Distributions

2.9.1 Poisson Distribution

$$X \sim P(\lambda = average)$$

PMF:

$$p(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, 2... \\ 0 & elsewhere \end{cases}$$

Models number of events, occurring randomly with an average rate of occurence per unit of time/space/distance

$$E(X) = Var(X) = \lambda$$

2.9.2 Exponential Distribution

$$X \sim E(\lambda = average/unit)$$

PDF:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & elsewhere \end{cases}$$

Models space/distance between events, occurring randomly, with an average rate of occurrence

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

2.10 Work Unit 4: Normal Distribution

Pattern of averages: If the random variable X is the sum of a large number of random increments then X has a normal distribution (Central Limit Theorem)

EG: Height of trees, amount of stuff in a jar
Continuous so PDF

$$X \sim N(\mu, \sigma^2)$$

$$p(x < \mu) = 0.5 \qquad p(x > \mu) = 0.5$$

PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Can't integrate analytically so use the tables in introstat. first convert to z:

$$z = \frac{x - \mu}{\sigma}$$

Tables give $P(0 < Z < z)$ so convert accordingly knowing that $P(Z < 0) = 0.5$ and the distribution is symmetric about 0.

Sum:

$$X_i \sim N(\mu_i, \sigma_i^2) \qquad Y = \sum X_i \qquad then \qquad Y \sim N(\sum \mu_i, \sum \sigma_i^2)$$

Difference:

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad X_2 \sim N(\mu_2, \sigma_2^2) \quad Y = X_1 - X_2$$

$$Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Multiplying by Constant:

$$X \sim N(\mu, \sigma^2) \qquad Y = aX + b \qquad Y \sim N(a\mu + b, a^2\sigma^2)$$

If you want to find $P(X > z^{(p)}) = p$, search for p or closest value in table. $z^{(0.1)}$ = upper 10%, lower 10% $= -z^{(0.1)} = z^{(0.9)}$ = upper 90%

2.11 Module 5: Hypothesis Testing

2.12 Work Unit 1: Sampling Distribution

Population

- Parameter - greek
- mean $= \mu$
- variance $= \sigma^2$

vs Sample

- Statistic - roman
- mean $= \bar{x}$
- variance $= s^2$
- statistic easier to measure
- can draw inference about population

Steps

1. Draw Random Sample
2. Measure sample statistic
3. use this as an estimate of the true unknown population parameter

Sample must be:

1. **Representative:** similar in structure to the population it is drawn from
2. **Random:** Every member of the population has an equal chance of being chosen as part of the sample

Statistics will vary due to random sample so a statistic is a random variable so $\bar{x}$ is a random variable with a probability distribution called a sampling distribution.

$\sum$ elements in a normal distribution has a normal distribution.
So for X_i drawn randomly from a normal distribution:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(\sum_{i=1}^{n} X_i) = n\mu$$

Multiplying by constant $^1/_n$:

$$E(\bar{X}) = \frac{1}{n}n\mu = \mu$$

Variance:

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

so:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

What if X_i is not normally distributed?
Central Limit Theorem: The average (or sum divided by n) of a large number of variables always has a normal distribution.
How Large? n = 30 is sufficient

so we have 3 means:

1. sample
2. probability distribution (expected value)
3. Population

Note sample mean is $\bar{X}$ because $\bar{x}$ is a specific value.

Sample mean based on a large n has a smaller variance so closer to μ

2.13 Work Unit 2: Confidence Intervals

Point Estimate: No information regarding the uncertainty of the estimate
vs **Interval:** range of values, communicate how much precision or uncertainty is present in the estimate

$$Pr(\bar{X} - z^{(\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z^{(\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha = 2L$$

we are $100(1-\alpha)\%$ confident that the value of μ lie in this interval choose $z^{(\frac{\alpha}{2})}$ from table such that $P(0 < z < z^{(\frac{\alpha}{2})}) = L$, given σ^2 from population

100(1-α)%	z$^{(\frac{\alpha}{2})}$
95%	1.96
90%	1.645
98%	2.33
99%	2.58

$$L = z^{(\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z^{(\frac{\alpha}{2})} \sigma}{L} \right)^2$$

width of confidence interval $= 2L$, if $\alpha = 0.05$ then confidence interval = 95%.

increase n, confidence interval narrows.
Increase z confidence interval widens

2.14 Work Unit 3: Testing whether the mean is a specific value

6 Step Hypothesis test

1. Define null hypothesis (H_0)

$$H_0 : \mu = 0$$

2. Define alternative hypothesis (H_1)

$$H_1 : \mu < a$$

   Or > (one-sided) or $\neq$ (2 sided)

3. Set significance level

$$\alpha = 0.05$$

   You will erroneously reject H_0 α% of the time

4. Set up rejection region
   Find $z^{\alpha/2}$ (2-sided) or z^{α} (one-sided)

5. Calculate the test statistic (Assume H_0 is true)

$$X \sim N(\mu_0, \frac{\sigma^2}{n})$$

$$z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

6. Draw a conclusion:
   If test statistic fall in rejection region reject H_0 in favour of H_1 else do not reject H_0

The default is a two-sided test unless you have good reason to suspect a departure in a specific direction.

Remember to split α in a two sided test

2.14.1 Errors

1. Type 1

   - Reject H_0 erroneously
   - controlled by α
   - α small reduces probability of this error
   - $P(T_1E) = \alpha$

2. Type 2

   - Accept H_0 erroneously
   - α small increases probability of this error
   - $P(T_2E)$ varies dependent on how close H_0 is to the true situation, so difficult to control

2.15 Work Unit 4: Comparing 2 Sample means

EG Test if 2 dies come from same or different populations

To compare look at difference:

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

Hypothesis Test

1. $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$
2. $H_1 : \mu_1 \neq \mu_2$
3. set α
4. Find rejection region (in this case 2 sided test but can be one-sided)
5. Calculate test statistic:

$$z = \frac{\bar{X_1} - \bar{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

2.15.1 The Modified Approach

Don't specify a significance level, instead observe significance level based on the test statistic.

Test Statistic:

- if H_1 is $\neq$ (2-sided): $p - value = P(z > |teststat|) \times 2$
- if H_1 is $>$ (1-sided): $p - value = P(z > TestStat)$
- if H_1 is $<$ (1-sided): $p - value = P(z < TestStat)$

If H_0 is true we would observe a difference of at least the size $X_1 - X_2$ p-value% of the time.

Small p-value means H_0 unlikely

reject H_0 if p-value< 0.05 (remember p-value is prob of type 1 error)

2.16 Work Unit 5: Tests about the mean when we don't know the variance

Estimate σ^2 from s^2

- Now two random variables $\bar{X}$ and s^2
- Test statistic now t-test
- still bell shaped and symmetric but flatter-fatter tails
- increase n looks more normal, smaller n - heavier tails
- t distribution
$$t = \frac{X - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$
has $n - 1$ degrees of freedom

In table: degrees of freedom along left column, % > along top

Confidence interval
$$\bar{X} \pm t_{n-1}^{\alpha/2} \frac{s}{\sqrt{n}}$$

Hypothesis test the same, just use t-table

Alternative Hypothesis test, for p-value look along n-1 row fro largest value that the test statistics exceeds

Why n-1 Degrees of freedom?

If given $\bar{x}$ and $x_1, ..., x_{n-1}$ can determine x_n, this is why sample variance is multiplied by $^1/_{n-1}$

General Rule: For each parameter we need to estimate (s^2) prior to evaluating the current parameter of interest $(\bar{X})$, we lose 1 degree of freedom

for $n > 30$: $s^2 \approx \sigma^2$

2.17 Work Unit 6: Comparing Means of 2 Independent Samples

1. Define null Hypothesis

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 - \mu_2 = 0$$

2. Define Alternative Hypothesis

$$H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 - \mu_2 \neq 0$$

3. Define significance level α

4. Rejection region

5. Test Statistic (t-test)

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   But this is wrong because it does not have a t-dist

6. so use pooled variance:

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

Assuming population variances are equal i.e. s_1^2 and s_2^2 are viewed as estimates of the same true variance

7. t becomes:

$$t = t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

8. conclusion:

$$|test\ statistic| > t_{n_1+n_2-2}^{\alpha/2}$$

Use closest degrees of freedom if the one you are looking for isn't in the table

Also can use modified approach

2.18 Work Unit 7: Comparing Means of 2 Dependent Samples

What matters is difference (change) so Before - After
Data is paired

Hypothesis Test

1. Define null Hypothesis

$$H_0 : \mu_B = \mu_A$$

$$H_0 : \mu_B - \mu_A = 0$$

2. Define Alternative Hypothesis

$$H_1 : \mu_B \neq \mu_A$$

$$H_0 : \mu_B - \mu_A \neq 0$$

   can be 2-sided

3. Define significance level α

4. Rejection region

5. Test Statistic

$$d = X_B - X_A$$

   degrees of freedom is n-1, where n is number of pairs

$$t = \frac{\bar{d} - \mu}{\frac{s_d}{\sqrt{n}}}$$

6. conclusion

if two-sided double p-value

Critical Point: paired data are not independent of each other - repeated measures (i.e. same people)

confidence interval

$$\bar{d} \pm t_{n-1}^{\alpha/2} \frac{s_d}{\sqrt{n}}$$

2.19 Work Unit 8: Testing whether data fits a specific distribution

Goodness of fit test: check what we observe in a sample with what we expect under a specific hypothesis

6 Step Approach

1. H_0 : X has some pdf/ pmt
2. H_1 : X has some other distribution (always 1 sided test, don't split α)
3. $\alpha = 0.05$
4. Chi-squared distribution
   - has degrees of freedom
   - skewed to the right
   - always positive
   - df = number of categories-num parameters we estimate-1
5. Test Statistic

$$D^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(O_i)^2}{E_i} - n$$

   k = number of comparisons, O = observed, E = expected value
6. conclusion

Modified approach: calculate p-value from table: find largest value which is smaller than the test statistic

Lose a degree of freedom for each parameter you estimate

Need an expected value of at least 5 in each category, otherwise collapse categories. e.g. Poisson dis either choose λ and test or get λ from data, in which case df = n-2

For normal μ and σ^2 are estimated so df = n-3

There is a mathematical relationship between the normal and chi-squared distributions

2.20 Work Unit 9: Testing for an association between 2 categoric variables

Table, there is an association between rows (e.g. gender) and columns (e.g. job level). Counts in cells, assume data is random and representative

6 Step Approach

1. H_0 : there is no association between rows and columns
2. H_1 there is an association
3. $\alpha = 0.05$
4. $D^2 > \chi_{df}^{2\alpha}$ assume H_0 is true. one sided, compare observed and expected values
   DF = ([no rows]-1)([no cols]-1)
5. Test Statistic - want E_i is the variables are independent. Remember
   $$P(A \cap B) = P(A)P(B)$$
   if A and B independent. so:
   $$E_{ij} = \frac{Row_i Total \times Col_j Total}{GrandTotal}$$
   now use:
   $$D^2 = \sum(\frac{O_i^2}{E_i}) - n$$
6. Conclusion

Modified approach: calculate p-value from table or excel

2.21 Work Unit 10: Testing for a predictive relationship between 2 Numeric Variables

Linear relationship between 2 quantitative random variables:

$$y = a + bx$$

y = dependent variable
x = independent variable
a + b = regression coefficients

Use correlation coefficient - true value unknown so estimate from sample:
ρ = population correlation (parameter)
r = Sample Correlation (statistic)

$$-1 \leq r \leq 1$$

r = -1 perfect negative (x inc, y dec), r = 1 perfect positive (x inc, y inc), r = 0 variables independent

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Coefficient of determination:

$$R^2 = r \times r \quad 0 \leq R^2 \leq 1$$

measure the property of variation in y that x is able to explain. $1-R^2$ is the proportion of variation in Y that is explained by factors other than X

$$y = \alpha + \beta x$$

α = y intercept
β = slope. estimate from sample
Closer —r— or R^2 to 1 the better the regression model fits the data, closer to 0, the worse the fit

Hypothesis Testing:

1. $H_0 : \beta = 0$ - no linear relationship
2. $H_1 : \beta \neq 0$ or $H_1 : \beta < 0$ or $H_1 : \beta > 0$
3. $\alpha = 0.01$
4. Rejection Region: test stat $\sim t_{n-2}$ where n = number of pairs of x and y (check other notes for tests)
5. Test statistic

3 Excel

1. =Rand()
2. =If(cond, value, or)
3. =countif(start:end, equals)