Originally by "'piccolojunior'" on the College Confidential forums; reformatted/reorganized/etc by Dillon Cower. Comments/suggestions/corrections: dcower@gmail.com

1

- Mean = $\bar{x}$ (**sample mean**) = μ (**population mean**) = sum of all elements ($\sum x$) divided by number of elements (n) in a set = $\frac{\sum x}{n}$. The **mean** is used for quantitative data. It is a measure of center.
- Median: Also a measure of center; better fits skewed data. To calculate, sort the data points and choose the middle value.
- Variance: For each value (x) in a set of data, take the difference between it and the mean ($x - \mu$ or $x - \bar{x}$), square that difference, and repeat for each value. Divide the final result by n (number of elements) if you want the **population variance** (σ^2), or divide by $n - 1$ for **sample variance** (s^2). Thus: Population variance = $\sigma^2 = \frac{\sum(x-\mu)^2}{n}$. Sample variance = $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$.
- Standard deviation, a measure of **spread**, is the square root of the variance. **Population standard deviation** = $\sqrt{\sigma^2} = \sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$. **Sample standard deviation** = $\sqrt{s^2} = s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$.
  - You can convert a population standard deviation to a sample one like so: $s = \frac{\sigma}{\sqrt{n}}$.
- Dotplots, stemplots: Good for small sets of data.
- Histograms: Good for larger sets and for **categorical** data.
- Shape of a distribution:
  - Skewed: If a distribution is skewed-left, it has fewer values to the left, and thus appears to **tail off** to the left; the opposite for a skewed-right distribution. If skewed right, **median** < **mean**. If skewed left, **median** > **mean**.
  - Symmetric: The distribution appears to be symmetrical.
  - Uniform: Looks like a flat line or perfect rectangle.
  - Bell-shaped: A type of symmetry representing a normal curve. Note: **No** data is perfectly normal - instead, say that the distribution is **approximately normal**.

2

- Z-score = standard score = normal score = z = number of standard deviations past the mean; used for **normal distributions**. A **negative z-score** means that it is below the mean, whereas a **positive z-score** means that it is above the mean. For a population, $z = \frac{x-\mu}{\sigma}$. For a sample (i.e. when a sample size is given), $z = \frac{x-\bar{x}}{s} = \frac{x-\bar{x}}{\frac{\sigma}{\sqrt{n}}}$.

- With a normal distribution, when we want to find the percentage of all values **less** than a certain value (x), we calculate x's z-score (z) and look it up in the Z-table. This is also the area under the normal curve to the left of x. Remember to multiply by 100 to get the actual percent. For example, look up $z = 1$ in the table; a value of roughly $p = 0.8413$ should be found. Multiply by $100 = (0.8413)(100) = 84.13\%$.
  - If we want the percentage of all values **greater** than x, then we take the complement of that $= 1 - p$.
- The area under the entire normal curve is **always** 1.

3

- Bivariate data: 2 variables.
  - Shape of the points (linear, etc.)
  - Strength: Closeness of fit or the correlation coefficient (r). **Strong**, **weak**, or **none**.
  - Whether the association is positive/negative, respectively.
- It probably isn't worth spending the time finding r by hand.
- Least-Squares Regression Line (LSRL): $\hat{y} = a + bX$. **(hat is important)**
- $r^2 =$ The percent of variation in y-values that can be explained by the LSRL, or how well the line fits the data.
- Residual = observed − predicted. This is basically how far away (positive or negative) the observed value (y) for a certain x is from the point on the LSRL for that x.
- **ALWAYS** read what they put on the axes so you don't get confused.
- If you see a pattern (non-random) in the residual points (think **residual scatterplot**), then it's safe to say that the LSRL doesn't fit the data.
- Outliers lie outside the overall pattern. **Influential points**, which significantly change the LSRL (slope and intercept), are outliers that deviate from the rest of the points in the x direction (as in, the x-value is an outlier).

4

- Exponential regression: $\hat{y} = ab^x$. (anything raised to x is exponential)
- Power regression: $\hat{y} = ax^b$.
- We **cannot** extrapolate (predict outside of the scatterplot's range) with these.
- Correlation **DOES NOT** imply causation. Just because San Franciscans tend to be liberal doesn't mean that living in San Francisco **causes** one to become a liberal.

- Lurking variables either show a **common response** or **confound**.
- Cause: x causes y, no lurking variables.
- Common response: The lurking variable affects both the explanatory (x) **and** response (y) variables. For example: When we want to find whether more hours of sleep explains higher GPAs, we must recognize that a student's courseload can affect his/her hours of sleep **and** GPA.
- Confounding: The lurking variable affects **only** the response (y).

5

- Studies: They're all studies, but observational ones don't impose a treatment whereas experiments do and thus we cannot do anything more than conclude a correlation or tendency (as in, NO CAUSATION)
- Observational studies do not impose a treatment.
- Experimental studies **do** impose a treatment.
- Some forms of bias:
  - Voluntary response: i.e. Letting volunteers call in.
  - Undercoverage: Not reaching all types of people because, for example, they don't have a telephone number for a survey.
  - Non-response: Questionnaires which allow for people to not respond.
  - Convenience sampling: Choosing a sample that is easy but likely non-random and thus biased.
- Simple Random Sample (SRS): A certain number of people are chosen from a population so that each person has an **equal** chance of being selected.
- Stratified Random Sampling: Break the population into strata (groups), then do a SRS on these strata. **DO NOT** confuse with a pure SRS, which does **NOT** break anything up.
- Cluster Sampling: Break the population up into clusters, then randomly select n clusters and poll all people in those clusters.
- In experiments, we **must** have:
  - Control/placebo (fake drug) group
  - Randomization of sample
  - Ability to replicate the experiment in similar conditions
- Double blind: Neither subject nor administrator of treatment knows which one is a placebo and which is the real drug being tested.
- Matched pairs: Refers to having each person do both treatments . Randomly select which half of the group does the treatments in a certain order. Have the other half do the treatments in the other order.

- Block design: Eliminate confounding due to race, gender, and other lurking variables by breaking the experimental group into groups (blocks) based on these categories, and compare only within each sub-group.
- Use a random number table or on your calculator: RandInt(lower bound #, upper bound #, how #'s to generate)

6

- Probabilities are ≥ 0 and ≤ 1.
- Complement $= 1 - P(A)$ and is written $P(A^c)$.
- Disjoint (aka **mutually exclusive**) probabilities have no common outcomes.
- Independent probabilities don't affect each other.
- $P(A\ and\ B) = P(A) * P(B)$
- $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$
- $P(B\ given\ A) = \frac{P(A\ and\ B)}{P(A)}$.
- $P(B\ given\ A) = P(B)$ means independence.

7

- Discrete random variable: Defined probabilities for certain values of x. Sum of probabilities **should** equal 1. Usually shown in a probability distribution table.
- Continuous random variable: Involves a density curve (area under it is 1), and you define intervals for certain probabilities and/or z-scores.
- Expected value = sum of the probability of each possible outcome times the outcome value (or payoff) $= P(x_1) * x_1 + P(x_2) * x_2 + \ldots + P(x_n) * x_n$.
- Variance $= \sum[(X_i - X_\mu)^2 * P(x_i)]$ for all values of x
- Standard deviation $= \sqrt{variance} = \sqrt{\sum(X_i - X_\mu)^2 P(x_i)}$
- Means of two different variables can add/subtract/multiply/divide. **Variances**, **NOT** standard deviations, can do the same. (Square standard deviation to get variance.)

8

- Binomial distribution: n is fixed, the probabilities of success and failure are constant, and each trial is independent.
- $p =$ probability of success
- $q =$ probability of failure $= 1 - p$
- Mean $= np$
- Standard deviation $= \sqrt{npq}$, which will only work if the mean (np) is ≥ 10 **and** nq ≥ 10.
- Use $binompdf(n, p, x)$ for a specific probability (exactly x successes).
- Use $binomcdf(n, p, x)$ sums up all probabilities up to x successes (including it as well). To restate this, it is the probability of getting x or fewer successes out of n trials.
  - The **c** in $binomcdf$ stands for **cumulative**.
- Geometric distributions: This distribution can answer two questions. Either a) the probability of getting first success on the nth trial, or b) the probability of getting success on $\leq n$ trials.
  - Probability of first having success on the nth trial $= p*q^{n-1}$. On the calculator: $geometpdf(p, n)$.
  - Probability of first having success on or before the nth trial = sum of the probability of having first success on the x trial for every value from 1 to $n = pq^0 + pq^1 + \ldots + pq^{n-1} = \sum_{i=1}^{n} pq^{i-1}$. On the calculator: $geometcdf(p, n)$.
  - Mean $= \frac{1}{p}$
  - Standard deviation $= \sqrt{\frac{q}{p^2}}$

9

- A statistic describes a **sample**. (s, s)
- A parameter describes a **population**. (p, p)
- $\hat{P}$ is a sample proportion whereas P is a parameter proportion.
- Some conditions:
  - Population size is ≥ 10 * sample size
  - np and nq must **both** be ≥ 10
- Variability = spread of data
- Bias = accuracy (closeness to true value)
- $\hat{P} =$ success/size of sample
- Mean $= \hat{p} =$ p
- Standard deviation: $\sqrt{\frac{pq}{n}}$

10

- H_0 is the null hypothesis
- H_a or H_1 is the alternative hypothesis.
- Confidence intervals follow the formula: estimator $\pm$ margin of error.
- To calculate a Z-interval: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
- The p value represents the chance that we should observe a value as extreme as what our sample gives us (i.e. how ordinary it is to see that value, so that it isn't simply attributed to randomness).
- If p-value is less than the alpha level (usually 0.05, but watch for what they specify), then the statistic is **statistically significant**, and thus we reject the null hypothesis.
- Type I error (α): We reject the null hypothesis when it's actually true.
- Type II error (β): We fail to reject (and thus accept) the null hypothesis when it is actually false.
- Power of the test $= 1 - \beta$, or our ability to reject the null hypothesis when it **is** false.

11

- T-distributions: These are very similar to Z-distributions and are typically used with small sample sizes or when the population standard deviation isn't known.
- To calculate a T-interval.
- Degrees of freedom (df) = sample size - 1 $= n - 1$
- To perform a hypothesis test with a T-distribution:
  - Calculate your test statistic: $t =$ (as written in the FRQ formulas packet) $\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$.
  - Either use the T-table provided (unless given, use a probability of .05 aka confidence level of 95%) or use the T-test on your calculator to get a t^* (critical t) value to compare against **your** t value.
  - If your t value is larger than t^*, then reject the null hypothesis.
  - You may also find the closest probability that fits your df and t value; if it is below 0.05 (or whatever), reject the null hypothesis.
- Be sure to check for normality first; some guidelines:
  - If $n < 15$, the sample must be normal with no outliers.
  - If $n > 15$ and $n < 40$, it must be normal with no outliers unless there is a strong skewness.
  - If $n > 40$, it's okay.
- Two-sample T-test:

- t = $\frac{\bar{x_1}-\bar{x_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.
- Use the smaller n out of the two sample sizes when calculating the df.
- Null hypothesis can be any of the following:
  * H_0: $\mu_1 = \mu_2$
  * H_0: $\mu_1 - \mu_2 = 0$
  * H_0: $\mu_2 - \mu_1 = 0$
- Use 2-SampTTest on your calculator.

- For two-sample T-test confidence intervals:
  - $\mu_1 \mu_2$ is estimated by $(\bar{x_1} - \bar{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - Use 2-SampTInt on your calculator.

12

- Remember ZAP TAX (Z for Probability, T for Samples ($\bar{X}$)).
- Confidence interval for two proportions:
  - $(\hat{p_1} - \hat{p_2}) \pm z^* \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}})$
  - Use 2-PropZInt on your calculator.
- Hypothesis test for two proportions:
  - $z = \frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1}+\frac{1}{n_2})}}$
  - Use 2-PropZTest on your calculator.
- Remember: Proportion is for categorical variables.

13

- Chi-square (χ^2):
  - Used for counted data.
  - Used when we want to test the independence, homogeneity, and "‘goodness of fit"’ to a distribution.
  - The formula is: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$.
  - Degrees of freedom = $(r-1)(c-1)$, where r = # rows and c = # columns.
  - To calculate the expected value for a cell from an observed table: $\frac{\text{(row total)(column total)}}{\text{table total}}$
  - Large χ^2 values are evidence against the null hypothesis, which states that the percentages of observed and expected match (as in, any differences are attributed to chance).

  - On your calculator: For independence/homogeneity, put the 2-way table in matrix A and perform a χ^2-Test. The expected values will go into whatever matrix they are specified to go in.

14

- Regression inference is the same thing as what we did earlier, just with us looking at the a and b in $\hat{y} = a + bx$.