Important facts from Introductory Statistics by G. Walther

Descriptive statistics

- Visualizations are helpful to communicate information and to support reasoning about the data.
- General design principles: provide context (compared to what?), avoid chart junk, avoid 3D plots.
- **Boxplots** give a concise summary (smallest number, 1st, 2nd, 3rd quartile, largest number) and can be used to compare data with the **principle of small multiples**.
- Numerical summaries of center: mean or median. Use the medican if the histogram is skewed as the mean can then be far away from the center.
- The spread can be measured by the standard deviation (SD) or the interquartile range (= 3rd quartile - 1st quartile).

Sampling

- **Population**: the entire group of subjects about which we want information; typically it's impossible to examine all subjects.
- **Parameter**: the quantity about the population which we are interested in.
- **Sample**: part of the population from which we collect information.
- **Statistic (estimate)**: the quantity we are interested in, but measured in the sample.
- Key point of statistical inference: The sample size determines the accuracy of the estimate, not the population size. So even a relatively small sample can produce an estimate that is close to the parameter, even if the population is very large.
- Sampling correctly is very important to avoid a bias. The best methods for sampling use chance in a planned way, such as the **simple random sample** (sample at random without replacement.)
- Sampling produces a chance error: estimate = parameter + bias + chance error.
- A larger sample size will typically reduce the chance error, but not the bias. We can compute the size of the chance error, but usually not the size of the bias.
- An **observational study** measures outcomes of interest and can be used to establish **association**. But **association is not causation**, because there may be **confounders**.
- A **randomized controlled experiment** can establish causation.

Probability

- Four basic rules: **Complement rule:** P(A does not occur)=1−P(A).
- **Rule for equally likely outcomes:** If there are n possible outcomes which are equally likely, then P(A)=(number of outcomes in A)/n.
- **Addition rule:** If A and B are mutually exclusive, then P(A or B)=P(A)+P(B).
- **Multiplication rule:** If A and B are independent, then P(A and B)=P(A)×P(B).
- **Bayes' rule:** $P(B|A)=P(A|B)\times P(B)/P(A)$, where the conditional probability $P(B|A)=P(A \text{ and } B)/P(A)$.

Sampling distributions and the central limit theorem

- Many data have histograms that follow the normal curve (are bell-shaped). Then the **empirical rule** applies: about 2/3 of the data are within one SD of the mean, and 95% are within 2 SDs.
- **Standardizing** data means subtracting off the mean, then dividing by the SD. The resulting standardized value is also called **z-score**; it has no units.
- The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.
- The **square root law** says that SE($\bar{x}_n$) decreases like $\frac{1}{\sqrt{n}}$.
- The **sampling distribution** of a statistic gives the probabilities of the outcomes of the statistic, e.g. it may follow the normal curve.
- There are three histograms which one has to keep carefully apart: The probability histogram that describes how the data are produced (typically unknown), the histogram of the observed data, the sampling distribution of a statistic computed from the data.
- The **law of large numbers** says that the sample mean $\bar{x}_n$ will be close to its expected value if the sample size is large.
- The **central limit theorem** says that the sampling distribution of $\bar{x}_n$ will be close to a normal curve when the sample size is large. So after standardizing (subtract off expected value, divide by SE) it will be close to a standard normal curve.

Regression

- A main task of statistical learning is to predict an outcome Y from an input X. Regression does that by fitting a linear function of X.
- **Regression effect** (**regression to the mean**): In a test-retest situation, the top group of the first test will drop down somewhat on the retest, while the bottom group moves up.
- There are checks whether regression is appropriate: The **residual plot** should not show a curved pattern nor be funnel-shaped (heteroscedastic).
- Transformations of X and/or Y may be necessary to make a linear model appropriate.
- **Outliers** should be scrutinized, **influential points** can invalidate conclusions.

Confidence intervals

- Confidence intervals give a range of plausible values for a parameter.
- Interpretation: Looking at many 95% confidence intervals, about 95% of them will trap the parameter and 5% will miss it.
- If the estimate is based on an average, then a simple formula for the confidence interval is: estimate $\pm$ zSE, where z is the z-score corresponding to the desired confidence level.
- Computing the SE of the estimate may involve other parameters that are unknown. The **bootstrap principle** says that one can replace those parameters with estimates from the sample and still get approximately correct confidence intervals.
- The **margin of error** is half the width of the confidence interval.

Testing hypotheses

- The **null hypothesis** states that nothing extraordinary is going on. We summarize the evidence in the data with a **test statistic** to decide whether there is sufficient evidence to reject the null hypothesis in favor of the **alternative hypothesis**.
- The most common test statistic is the **z-statistic = (observed - expected)/SE**. 'expected' and SE are computed under the assumption that the null is true. If the null is true, then the z-statistic will roughly follow a standard normal curve. Therefore this can serve as a reference to assess the evidence in the data: If the z-statistic is unusually large compared to a standard normal, then this is evidence to reject the null.
- The strength of the evidece is measured by the **p-value (observed significance level)**: it gives the probability of getting a value of z as extreme or more extreme than the observed z, assumning the null is true. So small values of the p-value (say $< 5\%$) can serve as a criterion to reject the null. Then the result is called **statistically significant**.
- The p-value does not give the probability that the null is true.
- If the sample size n is small (say $n < 20$) and the SE has to be estimated, then use a t-test with $n - 1$ degrees of freedom instead of the z-test.
- Statistically significant does not mean that the effect size is important! We may be strongly convinced that the null is false, even if the effect is small. Such a small effect may not be important to the investigator. For that reason it is helpful to complement the outcome of the test with a confidence interval for the parameter of interest.
- A **type I error (false positive)** occurs when the null is erroneously rejected. A **type II error** means that the test fails to reject a false null. Rejecting the null if the p-value is $< 5\%$ means P(type I error) $< 5\%$.
- A **two-sample z-test** compares the parameters of two populations by looking at the difference of two corresponding statistics. Care has to go into deriving the SE of that difference. It is applicable if the two samples are independent, and in special cases of dependent statistics such as in the case of a randomized controlled experiment.
- The **paired-difference test** is appropriate for analyzing the differences of **matched pairs** with a one-sample test.
- A simple version of the paired-difference test is the **sign test**, which simply counts how many differences are positive. Under the null, the outcome is like tossing a coin, so this test has the advantage of an easy interpretation.

Replicability and multiple testing

- If we perform many tests, then we will see a few statistically significant outcomes just by chance, even if there is no effect. Erroneously concluding that there is an effect is the **multiple testing fallacy**.
- Big data make it tempting to fall into this trap because there are so many things to explore in a large data set. This is called **data snooping** or **data dredging**.
- There are ways to account for data snooping: The **Bonferroni correction** divides the p-value by the number of tests conducted; it can be very conservative so that significant tests are missed.
- The **Benjamini-Hochberg procedure** is more powerful for discovering effects with multiple testing. It does this by focusing on the **false discovery rate** (the proportion of false discoveries among all discoveries) rather than the type I error.
- A simple way for exploring hypotheses is by holding out a **validation set** from this exploration. If an interesting hypothesis is found at his stage, it can then be tested in a second stage using the validation set.