9.1 Sample Statistics

- Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathsf{Var}[X_i] = \sigma^2$.
- The **sample mean** $\hat{\mu} = M_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ is often used to estimate the mean μ. It is an unbiased estimator with mean $\mathbb{E}[M_n] = \mu$ and variance $\mathsf{Var}[M_n] = \dfrac{\sigma^2}{n}$.
- The **sample variance** $\hat{\sigma}^2 = V_n = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} \big(X_i - M_n\big)^2$ is often used to estimate the variance σ^2. It is an unbiased estimator with mean $\mathbb{E}[V_n] = \sigma^2$.

9.2 New Families of Random Variables

- If $Z_1, \ldots, Z_n$ are i.i.d. Gaussian$(0,1)$ random variables, then $Y = \displaystyle\sum_{i=1}^{n} Z_i^2$ is a **chi-squared random variable with n degrees-of-freedom.**
  - Mean: $\mathbb{E}[Y] = n$
  - Variance: $\mathsf{Var}[Y] = 2n$
  - Shorthand Notation: $Y \sim \chi_n^2$
  - CDF: $\mathbb{P}[Y \leq y] = F_{\chi_n^2}(y)$ evaluated via lookup table or software. (`MATLAB: chi2cdf(y,n)`)
- If Z is a Gaussian$(0,1)$ random variable, Y is a chi-squared random variable with n degrees-of-freedom, and Y and Z are independent, then $W = Z\sqrt{\dfrac{n}{Y}}$ has a **Student's t-distribution with n degrees-of-freedom.**
  - Mean: $\mathbb{E}[W] = 0$ (for $n > 1$)
  - Variance: $\mathsf{Var}[W] = n/(n-2)$ (for $n > 2$)
  - Shorthand Notation: $W \sim T_n$
  - CDF: $\mathbb{P}[W \leq w] = F_{T_n}(w)$ evaluated via lookup table or software. (`MATLAB: tcdf(t,n)`)
  - PDF: Symmetric about 0. PDF onverges to a Gaussian$(0,1)$ PDF as n increases. $F_{T_n}(t) \approx \Phi(t)$ is a good approximation for $n \geq 30$.

9.3 Confidence Intervals

- Basic Idea: How can we estimate the mean from data and quantify the uncertainty in our estimate?
- Let $X_1, \ldots, X_n$ be i.i.d. random variables generated with a distribution with parameter θ (e.g., mean, variance). A **confidence interval** $[A, B]$ for the parameter θ with **confidence level** $1-\alpha$ satisfies $\mathbb{P}[A \leq \theta \leq B] = 1-\alpha$ where A and B are functions of $X_1, \ldots, X_n$.
- In practice, we usually see values such as $1-\alpha = 0.99,\ 0.95,\ 0.9$.
- Below, we develop confidence intervals for the mean that are often used in practice. The probability calculations are exact if we assume that $X_1, \ldots, X_n$ are i.i.d. Gaussian(μ, σ^2). For $n > 30$ samples, these are very good approximations if $X_1, \ldots, X_n$ are i.i.d. but not necessarily Gaussian.

9.3.1 Confidence Interval for the Mean: Known Variance

- When to use: Variance is known *or* $n > 30$ samples.

- Let $X_1, \dots, X_n$ be i.i.d. random variables with unknown mean μ and known variance σ^2. Then, $\big[M_n - \epsilon, M_n + \epsilon\big]$ with $\epsilon = \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)$ is a confidence interval for the mean μ with confidence level $1 - \alpha$.

- Recall that $Q(z)$ is the standard normal complementary CDF, $Q(z) = \Phi(-z) = 1 - \Phi(z)$.

- Intuition: The random interval $\big[M_n - \epsilon, M_n + \epsilon\big]$ captures the true mean with probability $1 - \alpha$. We use our prior knowledge of the variance to calculate this interval.

- If the variance σ^2 is unknown *and* we have $n > 30$ samples, we just substitute σ^2 with the sample variance V_n.

- Useful values: $Q^{-1}(0.05) = 1.64$, $Q^{-1}(0.025) = 1.96$, $Q^{-1}(0.005) = 2.57$

- `MATLAB:` $Q^{-1}(z) =$ `qfuncinv(z)`

9.3.2 Confidence Interval for the Mean: Unknown Variance

- When to use: Variance is unknown *and* $n \leq 30$ samples.

- Let $X_1, \dots, X_n$ be i.i.d. random variables with unknown mean μ and unknown variance σ^2. Then, $\big[M_n - \epsilon, M_n + \epsilon\big]$ with $\epsilon = -\frac{\sqrt{V_n}}{\sqrt{n}} F_{T_{n-1}}^{-1}\left(\frac{\alpha}{2}\right)$ is a confidence interval for the mean μ with confidence level $1 - \alpha$.

- Recall that $F_{T_{n-1}}(t)$ is the CDF for a Student's t-distribution with $n - 1$ degrees-of-freedom.

- Intuition: The random interval $\big[M_n - \epsilon, M_n + \epsilon\big]$ captures the true mean with probability $1 - \alpha$. We use the sample variance to calculate this interval.

- When $n > 30$, we should just use the known variance case, setting $\sigma^2 = V_n$, since the t-distribution is well-approximated by a Gaussian distribution in this regime.

- `MATLAB:` $F_{T_{n-1}}^{-1}(t) =$ `tinv(t,n-1)`

9.3.3 Confidence Interval for the Variance

- Let $X_1, \dots, X_n$ be i.i.d. random variables with unknown mean μ and unknown variance σ^2. Then, $\big[\beta_1 V_n,\ \beta_2 V_n\big]$ with $\beta_1 = \dfrac{n-1}{F_{\chi^2_{n-1}}^{-1}\left(1 - \frac{\alpha}{2}\right)}$ and $\beta_2 = \dfrac{n-1}{F_{\chi^2_{n-1}}^{-1}\left(\frac{\alpha}{2}\right)}$ is a confidence interval for the variance σ^2 with confidence level $1 - \alpha$.

- Intuition: The random interval $\big[\beta_1 V_n,\ \beta_2 V_n\big]$ captures the true variance with probability $1 - \alpha$.

- `MATLAB:` $F_{\chi^2_{n-1}}^{-1}(y) =$ `chi2inv(y,n-1)`

9.4 Significance Testing

- We only have a probability model for our observations under the **null hypothesis** H_0.
- The **significance level** $0 \leq \alpha \leq 1$ is used to determine when to **reject the null hypothesis**. Typical values: $\alpha = 0.01, 0.05, 0.1$.
- Given a **statistic** calculated from the dataset, the **p-value** is the probability of observing a value at least this extreme under the null hypothesis.
  - If $\mathrm{p-value} < \alpha$, then reject the null hypothesis.
  - If $\mathrm{p-value} \geq \alpha$, then fail to reject the null hypothesis.
- We will focus on significance tests for the mean:
  - A **one-sample test** compares the sample mean of a dataset to a baseline mean μ.
  - A **two-sample test** compares the sample means of two datasets to each other.
  - The probability calculations below are exact if we assume that the data is i.i.d. Gaussian under the null hypothesis. For $n > 30$ samples, the calculations are very good approximations if the data is are i.i.d. but not necessarily Gaussian under the null hypothesis.

9.4.1 One-Sample Z-Test

- Dataset: $X_1, \ldots, X_n$
- Null Hypothesis: Data is i.i.d. Gaussian(μ, σ^2) with known mean μ and known variance σ^2.
- Informally, does the mean of the data differ significantly from the baseline μ?
- **Procedure:**
  1. Calculate the sample mean $M_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
  2. Calculate the Z-statistic $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$.
  3. Calculate the $\mathrm{p-value} = 2\Phi(-|Z|)$ where $\Phi(z)$ is the standard normal CDF.
     MATLAB: $\Phi(z) =$ `normcdf(z)`
  4. If $\mathrm{p-value} < \alpha$, then reject the null hypothesis.
     If $\mathrm{p-value} \geq \alpha$, then fail to reject the null hypothesis.
- Useful values: $2\Phi(-1.64) = 0.1$, $2\Phi(-1.96) = 0.05$, $2\Phi(-2.57) = 0.01$
- In practice, it is reasonable to use this test when $n > 30$, even if the variance is estimated from data by the sample variance. In this regime, the Central Limit Theorem offers a good approximation.

9.4.2 One-Sample T-Test

- Dataset: $X_1, \ldots, X_n$
- Null Hypothesis: Data is i.i.d. Gaussian(μ, σ^2) with known mean μ and unknown variance σ^2.

- Informally, does the mean of the data differ significantly from the baseline μ?

- **Procedure:**

  1. Calculate the sample mean and sample variance

  $$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

  2. Calculate the Z-statistic $T = \dfrac{\sqrt{n}(M_n - \mu)}{\sqrt{V_n}}$.
  3. Calculate the $\mathrm{p-value} = 2F_{T_{n-1}}(-|T|)$ where $F_{T_{n-1}}(t)$ is the CDF for a Student's t-distribution with $n-1$ degrees-of-freedom.
     MATLAB: $F_{T_{n-1}}(t) =$ `tcdf(t,n-1)`
  4. If $\mathrm{p-value} < \alpha$, then reject the null hypothesis.
     If $\mathrm{p-value} \geq \alpha$, then fail to reject the null hypothesis.

- In practice, it is reasonable to use this test when $n \leq 30$, and the data is well-approximated by a Gaussian distribution.

9.4.3 Two-Sample Z-Test

- Dataset: $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$

- Null Hypothesis: $X_1, \ldots, X_{n_1}$ is i.i.d. Gaussian(μ, σ_1^2) and $Y_1, \ldots, Y_{n_2}$ is i.i.d. Gaussian(μ, σ_2^2) with known variances σ_1^2 and σ_2^2. The mean μ is unknown.

- Informally, do the datasets have the same mean?

- **Procedure:**

  1. Calculate the sample means $M_{n_1}^{(1)} = \dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} X_i$ and $M_{n_2}^{(2)} = \dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2} Y_i$.
  2. Calculate the Z-statistic $Z = \dfrac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.
  3. Calculate the $\mathrm{p-value} = 2\Phi(-|Z|)$ where $\Phi(z)$ is the standard normal CDF.
     MATLAB: $\Phi(z) =$ `normcdf(z)`
  4. If $\mathrm{p-value} < \alpha$, then reject the null hypothesis.
     If $\mathrm{p-value} \geq \alpha$, then fail to reject the null hypothesis.

- Useful values: $2\Phi(-1.64) = 0.1$, $2\Phi(-1.96) = 0.05$, $2\Phi(-2.57) = 0.01$

- In practice, it is reasonable to use this test when $n_1 > 30$ and $n_2 > 30$, even if the variances are estimated from data by the sample variances. In this regime, the Central Limit Theorem offers a good approximation.

9.4.4 Two-Sample T-Test

- Dataset: $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$
- Null Hypothesis: $X_1, \ldots, X_{n_1}$ is i.i.d. Gaussian(μ, σ^2) and $Y_1, \ldots, Y_{n_2}$ is i.i.d. Gaussian(μ, σ^2) with unknown, equal variance σ^2. The mean μ is unknown.
- Informally, do the datasets have the same mean?
- **Procedure:**
  1. Calculate the sample means and sample variances,
  $$M_{n_1}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \qquad M_{n_2}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$
  $$V_{n_1}^{(1)} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} \left(X_i - M_{n_1}^{(1)}\right)^2 \qquad V_{n_2}^{(2)} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} \left(Y_i - M_{n_2}^{(2)}\right)^2 ,$$
  and the pooled sample variance $\hat{\sigma}^2 = \dfrac{(n_1 - 1)V_{n_1}^{(1)} + (n_2 - 1)V_{n_2}^{(2)}}{n_1 + n_2 - 2}$.
  2. Calculate the T-statistic $T = \dfrac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$.
  3. Calculate the $\text{p} - \text{value} = 2F_{T_{n_1+n_2-2}}(-|T|)$ where $F_{T_{n_1+n_2-2}}(t)$ is the CDF for a Student's t-distribution with $n_1 + n_2 - 1$ degrees-of-freedom.
  MATLAB: $F_{T_{n_1+n_2-1}}(t) =$ `tcdf(t,n1+n2-1)`
  4. If $\text{p} - \text{value} < \alpha$, then reject the null hypothesis.
  If $\text{p} - \text{value} \geq \alpha$, then fail to reject the null hypothesis.
- In practice, it is reasonable to use this test when $n_1 \leq 30$ or $n_2 \leq 30$, and the data is well-approximated by a Gaussian distribution.
- For unknown, unequal variances, use Welch's T-test.